

# 30:3 BERKELEY TECHNOLOGY LAW JOURNAL

2015

Pages

1687

to

2132

Berkeley Technology Law Journal

Volume 30, Number 3

**Production:** Produced by members of the *Berkeley Technology Law Journal*.  
All editing and layout done using Microsoft Word.

**Printer:** Joe Christensen, Inc., Lincoln, Nebraska.  
Printed in the U.S.A.

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences—Permanence of Paper for Library Materials, ANSI Z39.48—1984.

**Copyright © 2015 Regents of the University of California.**  
All Rights Reserved.



Berkeley Technology Law Journal  
University of California  
School of Law  
3 Boalt Hall  
Berkeley, California 94720-7200  
btlj@law.berkeley.edu  
<http://www.btlj.org>

# BERKELEY TECHNOLOGY LAW JOURNAL

VOLUME 30

NUMBER 3

2015

## TABLE OF CONTENTS

### ARTICLES

PUBLIC HEALTH AS A MODEL FOR CYBERSECURITY INFORMATION SHARING.....	1687
<i>Elaine M. Sedenberg &amp; Deirdre K. Mulligan</i>	
CITIZEN SCIENCE: THE LAW AND ETHICS OF PUBLIC ACCESS TO MEDICAL BIG DATA .....	1741
<i>Sharona Hoffman</i>	
PRIVACY AND COURT RECORDS: AN EMPIRICAL STUDY.....	1807
<i>David S. Ardia &amp; Anne Klinefelter</i>	
PUSH, PULL, AND SPILL: A TRANSDISCIPLINARY CASE STUDY IN MUNICIPAL OPEN GOVERNMENT .....	1899
<i>Jan Whittington, Ryan Calo, Mike Simon, Jesse Woo, Meg Young &amp; Peter Schmiedeskamp</i>	
TOWARDS A MODERN APPROACH TO PRIVACY-AWARE GOVERNMENT DATA RELEASES .....	1967
<i>Micah Altman, Alexandra Wood, David R. O'Brien, Salil Vadhan &amp; Urs Gasser</i>	
OPEN DATA, PRIVACY, AND FAIR INFORMATION PRINCIPLES: TOWARDS A BALANCING FRAMEWORK.....	2073
<i>Frederik Zuiderveen Borgesius, Jonathan Gray &amp; Mireille van Eechoud</i>	

## SUBSCRIBER INFORMATION

The *Berkeley Technology Law Journal* (ISSN1086-3818), a continuation of the *High Technology Law Journal* effective Volume 11, is edited by the students of the University of California, Berkeley, School of Law (Boalt Hall) and is published in print three times each year (March, September, December), with a fourth issue published online only (July), by the Regents of the University of California, Berkeley. Periodicals Postage Rate Paid at Berkeley, CA 94704-9998, and at additional mailing offices. POSTMASTER: Send address changes to Journal Publications, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94720-7210.

**Correspondence.** Address all correspondence regarding subscriptions, address changes, claims for non-receipt, single copies, advertising, and permission to reprint to Journal Publications, University of California, Berkeley Law—Library, LL123 Boalt Hall—South Addition, Berkeley, CA 94705-7210; (510) 643-6600; [JournalPublications@law.berkeley.edu](mailto:JournalPublications@law.berkeley.edu). *Authors:* see section titled Information for Authors.

**Subscriptions.** Annual subscriptions are \$65.00 for individuals and \$85.00 for organizations. Single issues are \$30.00. Please allow two months for receipt of the first issue. Payment may be made by check, international money order, or credit card (MasterCard/Visa). Domestic claims for non-receipt of issues should be made within 90 days of the month of publication; overseas claims should be made within 180 days. Thereafter, the regular back issue rate (\$30.00) will be charged for replacement. Overseas delivery is not guaranteed.

**Form.** The text and citations in the *Journal* conform generally to the THE CHICAGO MANUAL OF STYLE (16th ed. 2010) and to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 20th ed. 2015). Please cite this issue of the *Berkeley Technology Law Journal* as 30 BERKELEY TECH. L.J. \_\_\_\_ (2015).

## BTLJ ONLINE

The full text and abstracts of many previously published *Berkeley Technology Law Journal* articles can be found at <http://www.btlj.org>. Our site also contains a cumulative index, general information about the *Journal*, the *Bolt*, a collection of short comments and updates about new developments in law and technology written by BTLJ members, and *BTLJ Commentaries*, an exclusively online publication for pieces that are especially time-sensitive and shorter than typical law review articles.

## INFORMATION FOR AUTHORS

The Editorial Board of the *Berkeley Technology Law Journal* invites the submission of unsolicited manuscripts. Submissions may include previously unpublished articles, essays, book reviews, case notes, or comments concerning any aspect of the relationship between technology and the law. If any portion of a manuscript has been previously published, the author should so indicate.

**Format.** Submissions are accepted in electronic format through the ExpressO online submission system. Authors should include a curriculum vitae and resume when submitting articles, including his or her full name, credentials, degrees earned, academic or professional affiliations, and citations to all previously published legal articles. The ExpressO submission website can be found at <http://law.bepress.com/expresso>.

**Citations.** All citations should conform to THE BLUEBOOK: A UNIFORM SYSTEM OF CITATION (Columbia Law Review Ass'n et al. eds., 20th ed. 2015).

**Copyrighted Material.** If a manuscript contains any copyrighted table, chart, graph, illustration, photograph, or more than eight lines of text, the author must obtain written permission from the copyright holder for use of the material.

# DONORS

The *Berkeley Technology Law Journal* and the Berkeley Center for Law & Technology acknowledge the following generous donors to Berkeley Law's Law and Technology Program:

## Partners

COOLEY LLP

FENWICK & WEST LLP

COVINGTON & BURLING LLP

ORRICK, HERRINGTON &  
SUTCLIFFE LLP

## Benefactors

FISH & RICHARDSON P.C.

SKADDEN, ARPS, SLATE,  
MEAGHER & FLOM LLP &  
AFFILIATES

KASOWITZ BENSON  
TORRES & FRIEDMAN LLP

WEIL, GOTSHAL & MANGES LLP

KIRKLAND & ELLIS LLP

WHITE & CASE LLP

LATHAM & WATKINS LLP

WILMER CUTLER PICKERING  
HALE AND DORR LLP

MCDERMOTT WILL & EMERY

WILSON SONSINI  
GOODRICH & ROSATI

MORRISON & FOERSTER LLP

WINSTON & STRAWN LLP

## Members

ALSTON & BIRD LLP

KEKER & VAN NEST LLP

BAKER BOTTS LLP

KILPATRICK TOWNSEND &  
STOCKTON LLP

BAKER & MCKENZIE LLP

KNOBBE MARTENS  
OLSON & BEAR LLP

BINGHAM MCCUTCHEN LLP

MUNGER, TOLLES & OLSON LLP

DURIE TANGRI LLP

O'MELVENY & MYERS LLP

GILLIN JACOBSON ELLIS  
LARSEN & LUCEY

PAUL HASTINGS LLP

GTC LAW GROUP LLP &  
AFFILIATES

ROPES & GRAY LLP

GUNDERSON DETTMER STOUGH  
VILLENEUVE FRANKLIN &  
HACHIGIAN, LLP

SIDLEY AUSTIN LLP

HAYNES AND BOONE, LLP

SIMPSON THACHER & BARTLETT  
LLP

HICKMAN PALERMO TRUONG  
BECKER BINGHAM WONG, LLP

TURNER BOYD LLP

HOGAN LOVELLS LLP

VAN PELT, YI & JAMES LLP

IRELL & MANELLA LLP

WEAVER AUSTIN VILLENEUVE &  
SAMPSON, LLP

# BOARD OF EDITORS

# 2015–2016

---

## *Executive Committee*

---

*Editor-in-Chief*  
JOSHUA D. FURMAN

*Managing Editor*  
MISHA TSUKERMAN

*Senior Articles Editors*  
RAVI ANTANI  
ZACHARY FLOOD  
KELLY VARGAS

*Senior Executive Editor*  
DIANA OBRADOVICH

*Senior Scholarship Editor*  
CAROLINA GARCIA

*Senior Annual Review Editors*  
GINNY SCHOLTES  
SORIN ZAHARIA

*Senior Online Content Editor*  
NOAH DRAKE

---

## *Editorial Board*

---

*Commentaries Editors*  
SWAROOP POUDEL  
YU TANEBE  
BONNIE WATSON

*Production Editors*  
ROXANA GUIDERO  
DUSTIN VANDENBERG

*Technical Editors*  
KRISTINA PHAM  
CHRISTOPHER YANDEL

*Annual Review Editors*  
CYNTHIA LEE  
BENJAMIN LI

*Notes & Comments Editors*  
WAQAS AKMAL  
ERICA FISHER

*Symposium Editors*  
TOMMY BARCZYK  
JOHN RUSSELL

*Submissions Editors*  
PHILIP MERKSAMER  
MAYA ZIV

*Web & Technology Editors*  
JOE CRAIG  
LEIGHANNA MIXTER

*External Relations Editor*  
SIMONE FRIEDLANDER

*Member Relations Editor*  
STEPHANIE CHENG

*Alumni Relations Editor*  
GOLDA CALONGE

*Web Content Editor*  
FAYE WHISTON

JESSICA ANNIS  
JOEL BROUSSARD  
CHRISTIAN CHESSMAN  
DANEILLE DEVLIN  
KELSEY GUANCIALE

*Articles Editors*  
YESOL HAN  
CASSY HAVENS  
MARK JOSEPH  
BILAL MALIK

CHRIS NORTON  
LIDA RAMSEY  
ERIC RIEDEL  
ARIEL ROGERS  
MAX SLADEK DE LA CAL

# MEMBERSHIP

Vol. 30 No. 3

---

*Associate Editors*

---

WILL BINKLEY	PAMUDH KARIYAWASAM	ALEJANDRO ROTHAMEL
BRITTANY BRUNS	HILARY KRASE	LUKAS SIMAS
DAPHNE CHEN	JOYCE LI	RICHARD SIMS
YUHAN (ALICE) CHI	MATTHEW MCCLELLAN	FERNANDA SOLIS CAMARA
KEVIN CHIU	GAVIN MOLER	SARAH SUWANDA
SARA CHUGH	CATALINA MONCADA	JON TANAKA
NOA DREYMAN	SHELBY NACINO	VALERIE TRUONG
RAN DUAN	JUAN NAZAR	RAOUL GRIFONI- WATERMAN
JORDAN FRABONI	ROBERT OLSEN	REID WHITAKER
JEREMY ISARD	EUNICE PARK	
NATASA JANEZ	JAIDEEP REDDY	

*Members*

---

YASMINE AGELIDIS	NOAH GUINEY	MICHELLE PARK
NIKI BAWA	BRIAN HALL	DYLAN PETERSON
KATE BRIDGE	ANDREA HALL	KEVIN PORMIR
JESSICA BRODSKY	JESSICA HOLLIS	CHRISTELLE PRIDE
KELSEA CARLSON	JENNIFER HSU	JING XUN QUEK
STEPHEN CHAO	KENSUKE INOUE	LEE REDFEARN
JOSEPH CHRISTIE	HYE JIN KIM	MATT RICE
RACHEL CORRIGAN	RITHIKA KULATHILA	FAITH SHAPIRO
MICHAEL DEAMER	SARAH KWON	DARINA
THOMAS DEC	TIFFANY LEUNG	SHTRAKHMAN
DARIUS DEGHAN	MEI LIU	JOSHUA STEELE
ELENA FALLOON	STASHA LOEZA	EVE TAI
MEGHAN FENZEL	SARAH MULLINS	MY THAN
EVAN FERGUSON	ELI NESS	CASEY TONG
WHITNEY FLORIAN	NATE NGEREBARA	ELISSA WALTER
ETHAN FRIEDMAN	PEGGY NI	MELISSA WEE
ANDREY GAVRILENKO	JESSICA OGLESBEE	TAMARA WIESEBRON
KAN GU	BARCLAY OUDERSLUYS	SHONG YIN
	ROBERT PARIS	YU ZHAO

# BTLJ ADVISORY BOARD

JIM DEMPSEY

*Executive Director of the  
Berkeley Center for Law & Technology  
U.C. Berkeley School of Law*

ROBERT C. BERRING, JR.

*Walter Perry Johnson Professor of Law  
U.C. Berkeley School of Law*

MATTHEW D. POWERS  
Tensegrity Law Group, LLP

JESSE H. CHOPER  
*Earl Warren Professor of Public Law  
U.C. Berkeley School of Law*

PAMELA SAMUELSON  
*Professor of Law & Information  
and Faculty Director of the  
Berkeley Center for Law & Technology  
U.C. Berkeley School of Law*

PETER S. MENELL  
*Professor of Law and Faculty  
Director of the Berkeley Center  
for Law & Technology  
U.C. Berkeley School of Law*

LIONEL S. SOBEL  
*Visiting Professor of Law  
U.C.L.A. School of Law*

ROBERT P. MERGES  
*Wilson Sonsini Goodrich & Rosati Professor  
of Law and Faculty  
Director of the Berkeley Center  
for Law & Technology  
U.C. Berkeley School of Law*

LARRY W. SONSINI  
Wilson Sonsini Goodrich & Rosati

REGIS MCKENNA  
*Chairman and CEO  
Regis McKenna, Inc.*

MICHAEL STERN  
Cooley LLP

DEIRDRE K. MULLIGAN  
*Assistant Professor and Faculty Director of  
the Berkeley Center for  
Law and Technology  
U.C. Berkeley School of Information*

MICHAEL TRAYNOR  
Cobalt LLP

JAMES POOLEY  
*Deputy Director General of the  
World Intellectual Property Organization*

THOMAS F. VILLENEUVE  
Gunderson Dettmer Stough Villeneuve  
Franklin & Hachigian LLP

BERKELEY CENTER FOR  
LAW & TECHNOLOGY  
2015–2016

---

*Executive Director*  
JIM DEMPSEY

*Faculty Directors*

KENNETH A. BAMBERGER	ROBERT P. MERGES	PAUL SCHWARTZ
PETER S. MENELL	DEIRDRE MULLIGAN	MOLLY S. VAN HOUWELING
	PAMELA SAMUELSON	

*Associate Director*  
LOUISE LEE

*Affiliated Faculty*

CHRIS JAY HOOFNAGLE	TALHA SYED	JENNIFER M. URBAN
---------------------	------------	-------------------

*In Memoriam*

SUZANNE SCOTCHMER  
1950–2014



# PUBLIC HEALTH AS A MODEL FOR CYBERSECURITY INFORMATION SHARING

*Elaine M. Sedenberg<sup>†</sup> & Deirdre K. Mulligan<sup>††</sup>*

## ABSTRACT

Policy proposals often feature information sharing as a means to improve cybersecurity, but lack specificity connecting these activities to specific goals intended to advance the state of cybersecurity. We use the Doctrine of Cybersecurity as a lens to examine existing information sharing efforts and evaluate the utility of information sharing proposals. Leaning on the analogous public good-oriented field of public health, we extract insights on how these information policies and practices evolved to promote goals while actively mediating among values. Based on our review of specific public health information sharing systems, we derive a set of four principles—expert and collaborative data governance, reporting minimization and decentralization, earliest feasible de-identification, and limitations on use—to guide the development of information sharing proposals within the cybersecurity context, and include an analysis of specific sharing mechanisms—data access modes and sharing platforms—that inform the implementation of these four principles. We conclude with a set of recommendations for consideration within the context of cybersecurity information sharing proposals.

---

DOI: <http://dx.doi.org/10.15779/Z38PZ61>

© 2015 Elaine M. Sedenberg & Deirdre K. Mulligan.

<sup>†</sup> Elaine Sedenberg is a Ph.D. student at the University of California, Berkeley, School of Information, where she is a National Science Foundation (NSF) Graduate Research Fellow and a Berkeley Graduate Fellow. Previously, Elaine was a Science Policy Fellow at the Science and Technology Policy Institute (STPI) in Washington D.C., which is a Federally Funded Research and Development Center that supports the White House Office of Science and Technology Policy (OSTP), as well as the NSF and other agencies.

<sup>††</sup> Deirdre K. Mulligan is an Associate Professor in the School of Information at UC Berkeley and a co-Director of the Berkeley Center for Law & Technology. Prior to joining the School of Information in 2008, she was a Clinical Professor of Law, founding Director of the Samuelson Law, Technology & Public Policy Clinic, and Director of Clinical Programs at the UC Berkeley School of Law (Boalt Hall). She is Chair of the Board of Directors of the Center for Democracy and Technology, and a Fellow at the Electronic Frontier Foundation. Prior to Berkeley, she served as staff counsel at the Center for Democracy & Technology in Washington, D.C.

## TABLE OF CONTENTS

I.	INTRODUCTION .....	1690
II.	INFORMATION SHARING THROUGH THE DOCTRINE OF PUBLIC CYBERSECURITY .....	1693
A.	CYBERSECURITY AS A PUBLIC GOOD .....	1693
B.	INFORMATION SHARING AS A MEANS TO ADVANCE PUBLIC CYBERSECURITY GOALS.....	1695
1.	<i>Information Sharing to Improve Artifacts, Policies, and Practices</i> .....	1696
a)	Sharing Information About Vulnerabilities .....	1696
b)	Sharing Information About Best Practices.....	1697
c)	Sharing Information About Threats and Risks.....	1699
2.	<i>Information Sharing to Manage and Respond to Vulnerabilities and Threats</i> .....	1700
III.	LEARNING FROM PUBLIC HEALTH .....	1702
A.	ESTABLISHING THE PUBLIC HEALTH AND CYBERSECURITY ANALOGY .....	1702
B.	THE ROLE OF INFORMATION SHARING IN PUBLIC HEALTH.....	1706
1.	<i>Prevention</i> .....	1707
2.	<i>Management and Response</i> .....	1708
IV.	PROMOTING GOALS AND MEDIATING AMONG VALUES: INSIGHTS FROM PUBLIC HEALTH .....	1711
A.	EXPERT AND COLLABORATIVE DECISIONS ABOUT DATA COLLECTION AND GOVERNANCE: PRACTICES, STANDARDS, AND RELEASE PROCEDURES .....	1713
B.	REPORTING MINIMIZATION AND DECENTRALIZATION.....	1717
C.	EARLIEST FEASIBLE DE-IDENTIFICATION .....	1719
D.	LIMITATIONS ON NON-PUBLIC HEALTH USES THAT NEGATIVELY AFFECT INDIVIDUAL INTERESTS.....	1721
E.	PUBLIC HEALTH INFORMATION SHARING MODELS.....	1724
1.	<i>Access to Federally Held Data</i> .....	1724
a)	Open Data: No Restrictions and Public Open Access .....	1725
b)	Limited Access with Some Restrictions.....	1727
c)	Internal Agency Use Only .....	1728
2.	<i>Non-Governmental Platforms for Public Health Information Sharing</i> .....	1729

V.	RECOMMENDATIONS FOR APPLYING INFORMATION SHARING LAWS AND POLICIES TO CYBERSECURITY .....	1730
A.	CLARIFY THE PUBLIC GOALS OF CYBERSECURITY AND THE ROLE OF INFORMATION SHARING IN ADVANCING THEM .....	1730
B.	CLARIFY CONNECTIONS BETWEEN DATA SHARING PROPOSALS AND PUBLIC GOALS.....	1730
C.	COORDINATE ACTIVITIES USING EXPERT COMMUNITIES .....	1731
D.	WHERE POSSIBLE, FOSTER VOLUNTARY INFORMATION SHARING COLLABORATIONS.....	1732
E.	EMPHASIZE DATA MINIMIZATION, DECENTRALIZATION, AND EARLY DE-IDENTIFICATION.....	1733
F.	PROVIDE ADDITIONAL PRIVACY PROTECTIONS THROUGH NATIONAL INFORMATION SHARING LAWS.....	1733
G.	MAKE AS MUCH CYBERSECURITY DATA AS POSSIBLE OPEN AND ACCESSIBLE FOR PUBLIC USE .....	1734
H.	CYBERSECURITY SHARING PRACTICES SHOULD EMPHASIZE ETHICAL PUBLIC CYBERSECURITY RESEARCH .....	1734
VI.	CONCLUSION.....	1736
VII.	APPENDIX.....	1737

## I. INTRODUCTION

Information sharing figures prominently in policy proposals to improve cybersecurity, yet the connection between information sharing—*a means*—and specific cybersecurity goals has not been clearly or convincingly argued. In response to cybersecurity incidents,<sup>1</sup> Congress<sup>2</sup> and the White House<sup>3</sup> have made various proposals to promote information sharing between private industry and the U.S. government. These proposed frameworks and passage of recent legislation<sup>4</sup> lack specificity about the data to be shared and governing practices to be employed.<sup>5</sup> They also fail to adequately address civil liberties issues or articulate the overarching goals and specific objectives information sharing will advance.<sup>6</sup>

The lack of clarity around goals operates at two levels. First, cybersecurity conversations lack a strong doctrinal foundation from which

---

1. Jose A. DelReal, *Eyes Turn to the Next Congress as Sony Hack Exposes Cybersecurity Flaws*, WASH. POST (Dec. 18, 2014), <http://www.washingtonpost.com/blogs/post-politics/wp/2014/12/18/eyes-turn-to-the-next-congress-as-sony-hack-exposes-cybersecurity-flaws/>; Information About OPM Cybersecurity Incidents, U.S. OFFICE OF PERS. MGMT. (July 17, 2015), <https://www.opm.gov/cybersecurity/>; Brian Krebs, *Posts Tagged: Target Data Breach*, KREBS ON SECURITY, <http://krebsonsecurity.com/tag/target-data-breach/> (last visited Oct. 17, 2015).

2. Cybersecurity Information Sharing Act, S. 754, 114th Cong. (2015), <https://www.congress.gov/114/bills/s754/BILLS-114s754pcs.pdf>.

3. Exec. Order No. 13,691, Promoting Private Sector Cybersecurity Information Sharing, 80 Fed. Reg. 9349 (Feb. 20, 2015), <https://www.archives.gov/federal-register/executive-orders/2015.html>; Exec. Order No. 13,636, Improving Critical Infrastructure Cybersecurity, 78 Fed. Reg. 11,739 (Feb. 19, 2013), <https://www.archives.gov/federal-register/executive-orders/2013.html>.

4. Consolidated Appropriation Act, 2016, Pub. L. No. 114-113, div. N, tit. I, <https://www.congress.gov/114/bills/hr2029/BILLS-114hr2029enr.pdf> (the “Cybersecurity Act of 2015”); see also *Congress Passes the Cybersecurity Act of 2015*, NAT’L L. REV. (Dec. 20, 2015), <http://www.natlawreview.com/article/congress-passes-cybersecurity-act-2015>.

5. See Jennifer Granick, *The Right Way to Share Information and Improve Cybersecurity*, JUST SEC. (Mar. 26, 2015, 10:53 AM), <http://justsecurity.org/21498/share-information-improve-cybersecurity> (arguing that none of the plans proposed by Congress or the White House “narrowly and specifically identifies the categories of information that Congress wants to allow to be shared”).

6. *Cyber-Surveillance Bill to Move Forward, Secretly*, CTR. FOR DEMOCRACY & TECH. (Mar. 4, 2015), <https://cdt.org/insight/cyber-surveillance-bill-to-move-forward-secretly/> (arguing that the Cybersecurity Information Sharing Act has moved “backwards in terms of privacy and civil liberties protections”); Mark Jaycox, *EFF to Congress: Stop the Cybersurveillance Bills*, EFF DEEPLINKS BLOG (Apr. 22, 2015), <https://www.eff.org/deeplinks/2015/04/eff-congress-stop-cybersurveillance-bills> (arguing that the Cybersecurity Information Sharing Act’s “vague definition,” as well as broad legal immunities for the government and companies, could lead to increased government surveillance and the sharing of information beyond the scope of cybersecurity objectives).

to evaluate proposed interventions. To what end is information sharing directed? Sharing information in and of itself will not improve cybersecurity. Policy proposals are motivated by a belief that information—that is currently unavailable to relevant parties—is necessary for certain cybersecurity-promoting activities. Yet, it is unclear exactly what activities policy makers want information sharing to fuel. What are recipients of information expected, or required, to do with information they receive? Whatever their private interest suggests? Or is there a broader shared set of public goals that should guide how recipients use this information? Is the goal of information sharing to aid law enforcement in identifying and prosecuting bad actors? Or is the goal to fuel vulnerability patching? Or is the hope that shared information will aid administrators in identifying and containing attacks in real time? Some combination of the three, or something else entirely? Clarifying the overarching goals of national cybersecurity policy is a precursor to a meaningful discussion about the likely effectiveness and relative appropriateness of sharing information.

Second, at the tactical level, current information sharing proposals do not specify the connections between the kinds of information to be shared and particular cybersecurity-promoting activities. Again, information may support activities that improve individual entities' security posture, or enable some broader vision of cybersecurity, or both, but current proposals fail to make these connections or direct activity toward specific ends. In this environment, information sharing is debated in the abstract with little attention to its role in an overall strategic national agenda, and with insufficient details to consider how access to specific information can support tactical activities that advance national priorities.

The lack of clear cybersecurity goals and nuanced tactical examination impedes the tough conversations about how to weigh and protect other values in our efforts to improve cybersecurity—including privacy, freedom of expression, innovation, and competition. In those cases where information sharing makes for sound policy, this lack of clear goals and tactics precludes the careful construction of laws and mechanisms to mediate tensions between cybersecurity goals and other values.

Our objective is to advance the policy deliberations about information sharing as a means to advance cybersecurity. We do so in two ways. First, we situate the consideration of information sharing within the broader understanding that cybersecurity is a public good. Second, drawing from the analogous area of public health, we offer a set of principles to guide policy makers in the construction of information sharing arrangements

that prioritize, mitigate, and manage tensions among public values, and between the public good and private interests.

Part II positions the conversation about information sharing within the context of a growing agreement that cybersecurity is both a national priority and a public good. We concur with those who argue that, due to a range of public goods failures, individual market choices under-produce cybersecurity and therefore the state must play a role in advancing cybersecurity. We use the Doctrine of Public Cybersecurity<sup>7</sup> to evaluate the utility of information sharing. Under that doctrine, the goals of cybersecurity policy are to produce more secure artifacts and systems, and to promote security protective behaviors and effective management of the ongoing vulnerabilities that emerge from a constantly changing threat landscape. Viewed through this lens, the question is how and under what conditions information sharing can advance these twin goals of improving the security of systems and managing residual insecurity. We briefly review existing information sharing activities in the cybersecurity area to examine their relationship to these goals.

Next, in Part III, we explore the rich and diverse information sharing policies and practices in the analogous field of public health, and consider the utility and limitations of these approaches in advancing public cybersecurity goals.

In Part IV, we review specific public health policies and practices around information sharing, paying particular attention to those that mitigate the impact of public health activities on other public values and private interests. First, we show how information sharing plays an essential role in specific prevention and response activities within public health, and is facilitated through diverse mechanisms that combine law, policy, and technical approaches to manage competing interests and values. Second, we derive a set of four principles from the public health information sharing ecosystem—expert and collaborative data governance, reporting minimization and decentralization, earliest feasible de-identification, and limitations on use—to guide the development and consideration of information sharing proposals in the cybersecurity context. We conclude Part IV with an analysis of specific sharing mechanisms—data access modes and sharing platforms—that have the potential to inform implementation of the four principles.

In Part V we use these four principles derived from public health to develop a set of recommendations to guide the consideration of

---

7. *See infra* Part II.

cybersecurity information sharing proposals. We recommend combining public-use practices, open data sets, and more limited information sharing regimes—coupled with limits on non-cybersecurity related uses of shared data—to advance public cybersecurity goals.

## II. INFORMATION SHARING THROUGH THE DOCTRINE OF PUBLIC CYBERSECURITY

Cybersecurity information sharing proposals should be evaluated based on their capacity to address public goods related failures that hamper the production of more secure systems, and limit the ability to identify and respond to ongoing security vulnerabilities. We adopt the Doctrine of Public Cybersecurity as our frame for considering the utility of information sharing generally, and briefly analyze existing cybersecurity information sharing activities through its lens.

### A. CYBERSECURITY AS A PUBLIC GOOD

Cybersecurity is an important domestic and international priority. Successful attacks on critical infrastructure,<sup>8</sup> strategic national assets,<sup>9</sup> personal information,<sup>10</sup> and corporate secrets<sup>11</sup> all stem from vulnerabilities in the interconnected socio-technical systems commonly referred to as the Internet. Such systems store personal and corporate secrets, help us connect and manage critical infrastructure, and form the communication and coordination backbone for the country.

The current state of cybersecurity is viewed as insufficient to protect the national, corporate, and personal activities entrusted to the Internet.<sup>12</sup>

---

8. See INDUS. CONTROL SYS. CYBER EMERGENCY RESPONSE TEAM (ICS-CERT), ICS-CERT MONITOR: SEPTEMBER 2014-FEBRUARY 2015, at 2 (2015), [https://ics-cert.us-cert.gov/sites/default/files/Monitors/ICS-CERT\\_Monitor\\_Sep2014-Feb2015.pdf](https://ics-cert.us-cert.gov/sites/default/files/Monitors/ICS-CERT_Monitor_Sep2014-Feb2015.pdf).

9. See Trevor Hughes, *Calif. Attacks Send Warning that Internet Lines are 'Basically Unsecured'*, USA TODAY, (July 1, 2015, 8:31 PM), <http://www.usatoday.com/story/tech/2015/07/01/california-internet-service-restored/29563899/>; ICS-CERT MONITOR, *supra* note 8.

10. See Krebs, *supra* note 1.

11. *Economic Espionage and Trade Secret Theft: Are Our Laws Adequate for Today's Threats?* Hearing Before the Subcomm. on Crime & Terrorism of the S. Comm. on the Judiciary, 113th Cong. (2014) (statement of Randall C. Coleman, Assistant Director, Counterintelligence Division, FBI), <https://www.fbi.gov/news/testimony/combating-economic-espionage-and-trade-secret-theft>.

12. See JASON, JSR-10-102, SCIENCE OF CYBER-SECURITY 9 (2010), <http://fas.org/irp/agency/dod/jason/cyber.pdf>; MINORITY STAFF OF HOMELAND SEC. AND GOVERNMENTAL AFFAIRS COMM., THE FEDERAL GOVERNMENT'S TRACK RECORD ON CYBERSECURITY AND CRITICAL INFRASTRUCTURE 2 (2014),

The failure of the market to produce adequate investments in information security is well-documented,<sup>13</sup> and explained by its public good qualities. Researchers have identified several public good characteristics that contribute to the chronic underproduction of cybersecurity.<sup>14</sup> Due to the network effects of security investments, individual actors are unable to reap the full value of their cybersecurity investments, or to limit their risk through independent investments.<sup>15</sup> Information asymmetry, combined with the misaligned incentives that these externalities cause, contribute to poor cybersecurity investments and management.<sup>16</sup> Further depressing investment in cybersecurity is the difficulty in assessing both risk and return on investment, which in turn creates difficulties for security professionals who must argue for dollars without strong metrics for success.

---

<http://www.hsgac.senate.gov/download/the-federal-governments-track-record-on-cybersecurity-and-critical-infrastructure> (prepared by Sen. Tom Coburn); DEP'T OF HOMELAND SEC., ENABLING DISTRIBUTED SECURITY IN CYBERSPACE: BUILDING A HEALTHY AND RESILIENT CYBER ECOSYSTEM WITH AUTOMATED COLLECTIVE ACTION 5 (2011), <http://www.dhs.gov/xlibrary/assets/nppd-cyber-ecosystem-white-paper-03-23-2011.pdf>.

13. Alessandro Acquisti, William Horne & Charles Palmer, *Cyber Economics*, in NATIONAL CYBER LEAP YEAR SUMMIT 2009 CO-CHAIRS' REPORT 25, 25 (2009), [https://www.nitrd.gov/nitrdgroups/images/b/bd/National\\_Cyber\\_Leap\\_jYear\\_Summit\\_2009\\_CoChairs\\_Report.pdf](https://www.nitrd.gov/nitrdgroups/images/b/bd/National_Cyber_Leap_jYear_Summit_2009_CoChairs_Report.pdf); BRENT R. ROWE & MICHAEL P. GALLAGHER, PRIVATE SECTOR CYBER SECURITY INVESTMENT STRATEGIES: AN EMPIRICAL ANALYSIS 2 (2006), <http://www.econinfosec.org/archive/weis2006/docs/18.pdf> (presented at the Fifth Workshop on the Economics of Information Security); Amitai Etzioni, *Cybersecurity in the Private Sector*, 28 ISSUES SCI. & TECH. 58, 59 (2011), [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2356955](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2356955).

14. The first malware, the Morris Worm in 1998, propagated at such a fast rate it infiltrated and compromised (often shutting down) computers across the Internet, including U.S. military sites. THOMAS K. CLANCY, COMPUTER CRIME AND DIGITAL EVIDENCE: MATERIALS AND CASES 500 (2011). There are many examples of botnets, that when left unpatched or uncontained, end up impacting government computers or contractors. See, e.g., Brian Krebs, *U.S. Government Takes Down Coreflood Botnet*, KREBS ON SECURITY (Apr. 11, 2014), <http://krebsonsecurity.com/2011/04/u-s-government-takes-down-coreflood-botnet>.

15. For example, Microsoft changed their security update policy to include pirated copies of Windows operating system because patching has a positive network effect on all Windows machines—legal or pirated alike. Ina Fried, *Piracy-Check Mandatory for Windows Add-Ons*, CNET (July 26, 2005), <http://www.cnet.com/news/piracy-check-mandatory-for-windows-add-ons/>; Lawrence M. Walsh, *Pirated Software Security: Patching Pirated Software*, TECHTARGET (Mar. 2004), <http://searchsecurity.techtarget.com/Pirated-software-security-Patching-pirated-software>; ROWE & GALLAGHER, *supra* note 13, at 2.

16. Etzioni, *supra* note 13, at 59; Esther Gal-Or & Anindya Ghose, *The Economic Incentives for Sharing Security Information*, 16 INF. SYST. RES. 186, 187 (2005).

Despite recognition of these public good related challenges, cybersecurity policy has not been oriented to address them. Historically, cybersecurity policy has—for the most part implicitly—been shaped by the goals of deterrence reflected in criminal laws, and by risk management principles reflected in process-oriented security standards.<sup>17</sup>

Prior work urged the adoption of the Doctrine of Public Cybersecurity to orient public policy and private sector activities toward addressing these public good related challenges.<sup>18</sup> This work argues that cybersecurity policy should aim to spur the production of more secure systems, security-promoting behaviors, and activities to manage and respond to ongoing insecurity. The Doctrine of Public Cybersecurity steers policy makers away from less fruitful orientations, such as the deterrence-oriented strategies reflected in current criminal law, which do little to encourage the production of cybersecurity or to manage cyber-insecurity. We believe this is the correct orientation for national cybersecurity policy.

We use the Doctrine of Public Cybersecurity to explore the utility of information sharing. Through this lens, information sharing is valuable when it supports the production of more secure systems and behaviors, and/or aids in the management of and response to ongoing vulnerabilities. To the extent they are deemed useful to advance these two goals, information sharing policies and technical mechanisms should be considered, but only where constructed with affordances and constraints that attend to other competing public and private values.

#### B. INFORMATION SHARING AS A MEANS TO ADVANCE PUBLIC CYBERSECURITY GOALS

Today's cybersecurity environment boasts a wide range of information sharing activities. Some, like industry specific Information Sharing and Analysis Centers (ISACs)<sup>19</sup> and the United States Computer Emergency Readiness Team (US-CERT),<sup>20</sup> are long standing and supported by the government to promote sharing between trusted communities or industry-specific partners, as well as the public. Information Sharing and Analysis

---

17. Deirdre K. Mulligan & Fred B. Schneider, *Doctrine for Cybersecurity*, 140 DAEDALUS 70 (2011).

18. *Id.*

19. *About Us*, NAT'L COUNCIL OF INFORMATION SHARING & ANALYSIS CTRS. (ISACs), <http://www.isaccouncil.org/aboutus.html> (last visited Aug. 12, 2015).

20. *About Us*, U.S. COMPUTER EMERGENCY READINESS TEAM (US-CERT), <https://www.us-cert.gov/about-us> (last visited Aug. 12, 2015).

Organizations (ISAOs)<sup>21</sup> were recently added to complement existing ISACs and offer an alternative organization outside of specified industries (e.g., region, sector, sub-sector, etc.). Other information sharing activities have arisen independently in response to specific threats either discrete or ongoing, and have largely been the product of private decisions by security practitioners and their employers. Some are aimed at improving specific products, while others focus on sharing best practices, or on identifying and managing attacks. We briefly examine some existing efforts to highlight their diversity and their relationships to public cybersecurity goals, and note some organizational shortcomings and opportunities for improvement.

1. *Information Sharing to Improve Artifacts, Policies, and Practices*

a) *Sharing Information About Vulnerabilities*

There is a rich information market (white hat and black hat) for the discovery and exchange of information about vulnerabilities and exploits.<sup>22</sup> Vulnerability rewards programs (VRPs), also known as “bug bounties,” incentivize the reporting of information to organizations (namely software vendors) so that they can create patches to prevent exploitation. These programs are designed to promote disclosure to those in the position to patch them since discovered—but unreported—vulnerabilities may be sold on the black market as zero-day exploits (exploitable software vulnerabilities unknown to the vendor). However, the effectiveness and value of these programs is debated since vulnerabilities often command a higher price on the black market,<sup>23</sup> and some argue the commercialization of vulnerability information limits the availability of data and knowledge within security research.<sup>24</sup> Not all vendors utilize VRPs, but those that do offer varying participation guidelines and incentive structures, often

---

21. Exec. Order No. 13,691, 80 Fed. Reg. 9349 (Feb. 13, 2015), <https://www.gpo.gov/fdsys/pkg/FR-2015-02-20/pdf/2015-03714.pdf>.

22. Serge Egelman, Cormac Herley & Paul C. van Oorschot, *Markets for Zero-Day Exploits: Ethics and Implications*, 2013 NEW SEC. PARADIGM WORKSHOP 41, 41 (2013); LILLIAN ABLON, MARTIN C. LIBICKI & ANGREA A. GOLAY, RAND CORP., *MARKETS FOR CYBERCRIME TOOLS AND STOLEN DATA: HACKER'S BAZAAR*, at ix (2014), [http://www.rand.org/content/dam/rand/pubs/research\\_reports/RR600/RR610/RAND\\_RR610.pdf](http://www.rand.org/content/dam/rand/pubs/research_reports/RR600/RR610/RAND_RR610.pdf).

23. Matthew Finifter, Devdatta Akhawe & David Wagner, *An Empirical Study of Vulnerability Rewards Programs*, 22 USENIX SECURITY SYMP. 273, 273 (2013), <https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/finifter>.

24. David McKinney, *Vulnerability Bazaar*, 5 IEEE SECURITY & PRIVACY 69, 69 (Dec. 12, 2007).

including both monetary rewards and recognition.<sup>25</sup> In some cases, third party security vendors will set up VRPs for companies that do not offer incentives to report vulnerability information. For instance, in 2007 VeriSign offered monetary rewards for exploits found in the newly released Windows Vista operating system since at the time Microsoft did not offer a VRP.<sup>26</sup>

Vulnerability reporting resulting in patches serves a robust preventative function. However, coordination challenges, and the lack of uniform policy regarding the public release of information about vulnerabilities, can detract from its utility. For instance, if a reported vulnerability impacts multiple vendors, it is challenging to coordinate and accommodate patch times for the many organizations that may also be competitors.<sup>27</sup> In addition, vendors' incentives to patch are not as straightforward as they might seem. Acting to patch vulnerabilities comes with economic tradeoffs for the affected company. When a vulnerability is made public—even where accompanied by a patch—a vendor risks facilitating more reverse engineering on its products, which makes its software potentially more vulnerable.<sup>28</sup> Patching can also be disruptive for end users, so even when companies issue patches, users may not apply them.<sup>29</sup>

#### b) Sharing Information About Best Practices

Regulatory models and government and private institutions facilitate sharing information about cybersecurity best practices, policies, and procedures. Regulatory models that formally adopt, or refer to, industry-

25. Finifter et al., *supra* note 23, at 273; Sharon Solomon, *11 Essential Bug Bounty Programs of 2015*, TRIPWIRE (Feb. 10, 2015), <http://www.tripwire.com/state-of-security/vulnerability-management/11-essential-bug-bounty-programs-of-2015/>.

26. Brad Stone, *A Lively Market, Legal and Not, for Software Bugs*, N.Y. TIMES (Jan. 30, 2007), <http://www.nytimes.com/2007/01/30/technology/30bugs.html>.

27. Hasan Cavusoglu et al., *Efficiency of Vulnerability Disclosure Mechanisms to Disseminate Vulnerability Knowledge*, 33 IEEE TRANSACTIONS ON SOFTWARE ENGINEERING 171, 171 (2007).

28. Jay Pil Choi, Chaim Fershtman & Neil Gandal, *Network Security: Vulnerabilities and Disclosure Policy*, 58 J. IND. ECON. 868, 868 (2010). There are also alternative ways to report information like bug tracking systems that automatically facilitate the information exchange between users and vendors by reporting glitches and bugs, though the effectiveness of these systems varies depending on the design. *Towards the Next Generation of Bug Tracking Systems*, 2008 IEEE SYMP. VISUAL LANGUAGES & HUMAN-CENTRIC COMPUTING 82 (2008).

29. For instance, a recent study of the Heartbleed vulnerability noted that the number of patches deployed plateaued after two weeks, and that 3% of the Alexa Top One Million websites were still vulnerable two months after the disclosure. Zakir Durumeric et al., *The Matter of Heartbleed*, 2014 INTERNET MEASUREMENT CONF. 475, 475, <http://conferences.sigcomm.org/imc/2014/papers/p475.pdf>.

generated security standards indirectly encourage information sharing about security practices. One study suggests that involving private entities in the rule-making process through regulatory delegation models may have some positive impact on security outcomes.<sup>30</sup> This positive impact may be a result of the increased information sharing among companies promoted by the standard development process.

Incident response organizations designed to coordinate action or facilitate a response to a security compromise also advise entities on recommended security practices to reduce cyber vulnerability. The first CERT center was established in 1988 at Carnegie Mellon University (CMU).<sup>31</sup> US-CERT works with a spectrum of partners (e.g., from academia, industry, ISACs, security vendors, and state, local, or federal governments) and disseminates relevant threats and vulnerability information to targeted parties both large and small—from government, private sector, and the general public—in addition to their role in response management and coordination discussed below.<sup>32</sup> US-CERT publishes “Recommended Practices” to its website to encourage early implementation of known practices and configurations that would reduce the potential for an attack.<sup>33</sup> Additional CERTs operate internationally to provide complementary services, including setting standards, best practices, and policies across the world.<sup>34</sup> There are other federal efforts focused on improving artifacts, policies, and information sharing practices, including InfraGard<sup>35</sup> and the Secret Service Electronic Crimes Task

---

30. David Thaw, *The Efficacy of Cybersecurity Regulation*, 30 GA. ST. U. L. REV. 287 (2014).

31. The original CERT at Carnegie Mellon University is now referred to as CERT Coordination Center (CERT/CC) and works closely with US-CERT, which was established in 2003 and is a part of the Department of Homeland Security (DHS) National Cybersecurity and Communications Integration Center (NCCIC). Stuart Madnick, Xitong Li & Nazli Choucri, *Experiences and Challenges with Using CERT Data to Analyze International Cyber Security* 2, 5 (MIT Sloan Sch. of Mgmt. Working Paper CISEL #13, 2009), [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1478206](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1478206).

32. Gregory B. White & D.J. DiCenso, *Information Sharing Needs for National Security*, 38 HAW. INT’L CONF. ON SYS. SCI. 1, 5 (2005).

33. *Recommended Practices*, INDUS. CONTROL SYS. CYBER EMERGENCY RESPONSE TEAM (ICS-CERT), <https://ics-cert.us-cert.gov/Recommended-Practices> (last visited Aug. 23, 2015).

34. Madnick, Li & Choucri, *supra* note 31, at 2.

35. InfraGard is a nonprofit organization and public/private partnership between the FBI and private sector to facilitate the exchange of information in order to prevent hostile threats against the United States. INFRAGARD, <https://www.infragard.org> (last visited July 20, 2015).

Forces (ECTF).<sup>36</sup> These organizations, like the CERTs, share best practices in addition to information regarding emerging threats and existing risk.

c) Sharing Information About Threats and Risks

Information sharing collaborations between industry partners are another vital way to share knowledge about threats and risks as well as preventative measures. In 1998, Presidential Decision Directive 63 (PDD-63) identified distinct industries and called for the private sector within these industries to set up ISACs to share information to mitigate risk and promote effective responses to adverse events, including cyber events.<sup>37</sup> For example, there is an Information Technology ISAC (IT-ISAC)<sup>38</sup> and a Financial Services ISAC (FS-ISAC).<sup>39</sup> Organizing around industry sectors facilitates more specific information exchanges about vulnerabilities, threats, and isolated incidents. There are potential risks associated with exchanging security information, such as loss of competitive advantage, market share, or stock market value from negative publicity if information is inadvertently shared with competitors or the public. However, members benefit from industry-specific information exchanges that assist in prevention efforts and vulnerability identification and management.<sup>40</sup>

There are several privately organized cybersecurity threat information sharing platforms (often between market competitors) like McAfee's Cyber Threat Alliance<sup>41</sup> and Facebook's ThreatExchange.<sup>42</sup> There, the information exchange may not be explicitly linked to collective action

36. The Secret Service ECTF was created to form a network of diverse stakeholders (law enforcement, prosecutors, private industry, academics, etc.) for prevention, detection, mitigation, and investigation of cyber incidents. *Investigation*, SECRET SERVICE, <http://www.secretservice.gov/investigation> (last visited Oct. 18, 2015).

37. Presidential Decision Directive 63 on Critical Infrastructure Protection, 63 Fed. Reg. 41804-01 (Aug. 5, 1998); see also Daniel B. Prieto, *Information Sharing with the Private Sector: History, Challenges, Innovation, and Prospects*, in SEEDS OF DISASTER, ROOTS OF RESPONSE: HOW PRIVATE ACTION CAN REDUCE PUBLIC VULNERABILITY 404, 406 (July 14, 2006).

38. *Member ISACs*, NAT'L COUNCIL OF ISACs, <http://www.isaccouncil.org/memberisacs.html> (last visited Oct. 18, 2015).

39. FINANCIAL SERVICES INFORMATION SHARING AND ANALYSIS CENTER, <https://www.fsisac.com> (last visited Oct. 18, 2015).

40. Gal-Or & Ghose, *supra* note 16, at 187.

41. See Vincent Weafer, *McAfee Founds Cyber Threat Alliance with Industry Partners*, MCAFEE LABS (Sept. 29, 2014), <https://blogs.mcafee.com/mcafee-labs/mcafee-founds-cyber-threat-alliance-industry-partners>.

42. See Cade Metz, *Facebook Unveils Tool for Sharing Data on Malicious Botnets*, WIRED (Feb. 11, 2015), <http://www.wired.com/2015/02/facebook-unveils-tool-sharing-data-malicious-botnets>.

against threats and may only pertain to a narrow type of threat information—like spam propagation on spoiled URLs in ThreatExchange’s case.<sup>43</sup> Though these alliances often consist of only a few industry members, sharing information against common threats mutually increases the value and security of their respective security software and social media platforms.

Information sharing about emerging and existing threats and risks within an industry, particularly before they have been successfully exploited, can bolster prevention-related activities. Once these vulnerabilities are exploited, information sharing assists in coordinating action to manage the resulting insecurity.

2. *Information Sharing to Manage and Respond to Vulnerabilities and Threats*

Cybersecurity must manage residual insecurity by identifying both known and unknown threats and quickly mobilizing a response that contains and treats infected systems. In addition to improving artifacts, policies, and practices in order to further preventative cybersecurity measures, US CERT Coordination Center (CERT/CC),<sup>44</sup> ISACs, and Secret Service ECTFs all play roles in managing the exchange of information necessary to coordinate responses to cyber incidents.<sup>45</sup>

There are notable cases within cybersecurity where ad hoc groups of researchers coalesced to respond to an emerging threat. Technical ad hoc working groups of researchers and practitioners self-organized to respond to the 2008 emergence of an aggressive worm (dubbed “Conficker”) intended to create a botnet. Individuals from Microsoft, ICANN, domain registry operators, anti-virus vendors, and academic security researchers spontaneously formed the Conficker Working Group (CWG) to contain its spread and effectiveness.<sup>46</sup> Other botnet working groups, like the DNS Changer Working Group (DCWG) and the 2010 Mariposa working group, have followed the example of CWG, facilitating coordination and

---

43. *ThreatExchange*, FACEBOOK, <https://developers.facebook.com/products/threat-exchange> (last visited Aug. 23, 2015).

44. US-CERT, *supra* note 20.

45. Lawrence A. Gordon, Martin P. Loeb & William Lucyshyn, *Sharing Information on Computer Systems Security: An Economic Analysis*, 22 J. ACCT. & PUB. POL’Y 461, 463 (2003).

46. THE RENDON GROUP, CONFICKER WORKING GROUP: LESSONS LEARNED CONTRACT, at ii (2010), [http://www.confickerworkinggroup.org/wiki/uploads/Conficker\\_Working\\_Group\\_Lessons\\_Learned\\_17\\_June\\_2010\\_final.pdf](http://www.confickerworkinggroup.org/wiki/uploads/Conficker_Working_Group_Lessons_Learned_17_June_2010_final.pdf).

information exchange.<sup>47</sup> Similarly, Microsoft initiated a working group in 2012 led by the Microsoft Digital Crimes Unit (with additional support from Microsoft's Malware Protection Center, the FS-ISAC, and Electronic Payments Association) to orchestrate the seizure of the Zeus botnet.<sup>48</sup> These working groups coordinate to varying degrees with law enforcement.<sup>49</sup>

Though these working groups have succeeded, by some measure, in managing coordinated responses to threats, the Conficker retrospective report recommends “improve[d] cooperation between the private sector and the U.S. government and governments around the world so that information sharing and efforts become a two-way exchange” to improve future outcomes.<sup>50</sup> The report also calls for clarification on the private sector's relationship with law enforcement, and procedures for reporting early warning signals to the government.<sup>51</sup>

Researchers have noted other weaknesses in the cybersecurity information sharing landscape including difficulty obtaining data in a timely and consistent format,<sup>52</sup> organizational and policy challenges associated with the dissemination of vulnerability disclosures,<sup>53</sup> and inattention to the privacy risks associated with sharing relevant data.<sup>54</sup>

Current information sharing activities attest to the various ways information can address public good challenges to cybersecurity. Today, the exchange of various kinds of information supports system improvement, shared understandings of risks, coordinated action, and priority setting. But these efforts have arisen outside a comprehensive

47. Andreas Schmidt, *Hierarchies in Networks: Emerging Hybrids of Networks and Hierarchies for Producing Internet Security*, in CYBERSPACE AND INTERNATIONAL RELATIONS 181, 190–91 (Jan-Frederik Kremer & Benedikt Müller eds., 2013), <http://link.springer.com/10.1007/978-3-642-37481-4>.

48. *Id.* at 191.

49. *Id.* at 190; Milton Mueller, Andreas Schmidt & Brenden Kuerbis, *Internet Security and Networked Governance in International Relations*, 15 INT'L STUD. REV. 86, 96 (2013) (“In cases like CWG, law enforcement, intelligence agencies, and entities like CERT played a negligible role in containment. This further underscores the need for a cohesive cybersecurity response management strategy.”).

50. THE RENDON GROUP, *supra* note 46, at iv.

51. *Id.*

52. See Oscar Serrano, Luc Dandurand & Sarah Brown, *On the Design of a Cyber Security Data Sharing System*, 2014 ACM WORKSHOP ON INFO. SHARING COLLABORATION SECURITY 61, 61.

53. See Jennifer Granick, *The Price of Restricting Vulnerability Publications*, 9 INT'L J. COMM. L. & POL'Y (SPECIAL ISSUE) 1, 5 (2005).

54. See Gina Fisk et al., *Privacy Principles for Sharing Cyber Security Data*, 2015 IEEE INT'L WORKSHOP ON PRIVACY ENGINEERING 1, 1 (2015).

framework—such as the Doctrine for Public Cybersecurity—and thus appears as a set of loosely aligned activities rather than an integral component of a strategic, cohesive agenda. The somewhat ad hoc development of the information sharing environment, and mixture of public and private actors, has limited systematic, public consideration of their independent and collective impact on other public values. This in turn has fueled public concern about the impact of information sharing on privacy, freedom of expression, and human rights.

### III. LEARNING FROM PUBLIC HEALTH

The current cybersecurity information sharing ecosystem supports activities aligned with public cybersecurity goals. Yet, these activities have emerged in a piecemeal and sporadic manner, lacking a strong vision of the potential role information sharing could play in advancing public priorities and a framework to ameliorate their impact on other values. Information sharing activities in the more mature public health domain, which address similar public goods challenges, offer insight into how a developed and coordinated information sharing system between diverse stakeholders can advance public cybersecurity goals and protect other public values.

Information sharing is pervasive in the field of public health. It plays an essential role in promoting two key public health goals: advancing the health of the population by addressing the fundamental causes of disease; and preventing adverse health outcomes in a manner that enhances the physical and social environment while respecting the rights of individuals.<sup>55</sup> Furthermore, the public health field has developed policies and mechanisms to balance competing values that arise in information sharing activities.

Below we discuss differences and similarities between the public health and cybersecurity domains that could affect the utility of similar activities in the realm of cybersecurity, and examine information sharing activities and their role in advancing public health functions within the field.

#### A. ESTABLISHING THE PUBLIC HEALTH AND CYBERSECURITY ANALOGY

The Doctrine of Public Cybersecurity differentiates between population level goals and individual responses to threats and security

---

55. See PUBLIC HEALTH LEADERSHIP SOCIETY, PRINCIPLES OF THE ETHICAL PRACTICE OF PUBLIC HEALTH: VERSION 2.2 (2002), <http://phls.org/CMSuploads/Principles-of-the-Ethical-Practice-of-PH-Version-2.2-68496.pdf>.

incidents, understanding that due to its public goods characteristics, the interests of individuals, firms, and investors are not aligned to deliver an acceptable level of cybersecurity. The analogous field of public health responds to a similar problem. Public health focuses efforts at the population level—aiming to improve the functioning and longevity of the population by addressing underlying health issues and causes. Public health initiatives focus primarily on population-level responses to health concerns rather than the course of treatment for any one individual’s care. While individual health choices may advance the overall health of the population, at times the interests of the population and those of the individual are misaligned, or even at odds. In such instances, the government steps in to prompt or take actions that support the overall well-being of the population. Public health has developed mechanisms for balancing the tension between individual and collective interests in such instances.

Some have expressed reservations about the analogy between cybersecurity and public health.<sup>56</sup> Reservations include the prominence of the intelligent sentient adversary in cybersecurity, the complexity and severity of the tradeoffs given the expressive nature of some of the information subject to sharing, the large role of the private sector given its ownership and control of relevant infrastructure and possession of relevant data, and the mixed motives of the government, which has both a defensive and an offensive interest in cybersecurity.<sup>57</sup> Acknowledging that there are certainly limitations to the analogy, we believe they are more in quality than in kind. For example, while pathogens may not be intelligently adversarial—as we narrowly define intelligence—biologically they evolve in form and function to adapt to environmental changes or take advantage of changing social structures (e.g., rapid spreading through urbanization, or growing antibiotic resistance from prescription overuse).<sup>58</sup>

---

56. For example, consider the audio from the Q&A from the symposium presentation. *Panel 3: Comparative Approaches: Privacy Law and Public Health Law*, 19TH ANN. BCLT/BTLJ SYMP. (2015), <https://www.law.berkeley.edu/centers/berkeley-center-for-law-technology/past-events/april-2015-the-19th-annual-bcltblj-symposium-open-data-addressing-privacy-security-and-civil-rights-challenges/program>.

57. *Id.*

58. Scholars have used biological pathogens as a comparative model for phenomenon within cyberspace since the early days of networked computers. Fred Cohen used the term “computer virus” in 1983 to describe the spread and replication of malicious code in the laboratory and wild. *History of Viruses*, NAT’L INST. OF STANDARDS & TECH. COMPUTER SECURITY RESOURCE CTR. (Mar. 10, 1994), [http://csrc.nist.gov/publications/nistir/threats/subsubsection3\\_3\\_1\\_1.html](http://csrc.nist.gov/publications/nistir/threats/subsubsection3_3_1_1.html); FRED COHEN, *COMPUTER VIRUSES—THEORY AND EXPERIMENTS* (1984), <http://web.eecs.umich>

While biologically designed to survive, rather than a desire to maximize damage, the evolution that results yields an arms race that is a hallmark feature of cybersecurity. Further many cybersecurity threats like the perfunctory propagation of malware lacks sentience once released into the wild much like many biological threats. Thus while some motives vary, in both domains the public good is subject to a constantly changing battlefield of new vulnerabilities, new exploits, and wily, motivated adversaries.

More importantly, the focus on the adversarial difference often blinds us to the fact that preventative techniques are effective regardless of motive. Preventative techniques reduce vulnerabilities regardless of adversarial goals. This can be seen in public health examples such as condom use. While the intentional spreading of disease is relatively uncommon, there are instances where individuals have knowingly exposed others to HIV.<sup>59</sup> This intent to infect is atypical in public health, but condoms are an effective preventative measure regardless of the host's intent. Similarly, crimeware, which can be purchased en masse on the black market, is only successful if there are unpatched vulnerabilities in web applications, or if users download attachments in suspicious emails. Preventative techniques that patch vulnerabilities or limit downloads and executables are effective against exploits regardless of the enhanced abilities resulting from automation coupled with malicious intention. Adversarial considerations are simply less relevant when dealing with prevention and management orientations—in contrast to deterrence-

---

.edu/~aprakash/eecs588/handouts/cohen-viruses.html. Consequently words like hosts, infection, and network health have entered the cybersecurity lexicon. This biological analogy has been extended to distributed security in cyberspace, comparing its diversity to the natural ecosystem and complex system responses to the human immune system. See U.S. DEP'T OF HOMELAND SECURITY, ENABLING DISTRIBUTED SECURITY IN CYBERSPACE: BUILDING A HEALTHY AND RESILIENT CYBER ECOSYSTEM WITH AUTOMATED COLLECTIVE ACTION 8 (2011), <https://www.dhs.gov/xlibrary/assets/nppd-cyber-ecosystem-white-paper-03-23-2011.pdf>. Given the precedent of using biological pathogens and immune system defenses as a comparative model for cybersecurity, the comparison between public health and public cybersecurity as public goods is particularly apt.

59. See Mary D. Fan, *Sex, Privacy and Public Health in a Casual Encounters Culture*, 45 U.C. DAVIS L. REV. 431 (2011). Prosecution in the United States differs by jurisdiction, and different states criminalize different behaviors specific to the knowing transmission or exposure of HIV. Philip B. Berger, *Prosecuting for Knowingly Transmitting HIV is Warranted*, 180 CAN. MED. ASS'N J. 1368 (2009), <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2696543>; Michael E. Miller, *Man Who Knowingly Spread HIV Sentenced to Six Months. Judge Calls it a 'Travesty'*, WASH. POST (May 5, 2015), <http://www.washingtonpost.com/news/morning-mix/wp/2015/05/05/man-who-knowingly-spread-hiv-sentenced-to-six-months-judge-calls-it-a-travesty>.

oriented strategies that are focused on intent—because harms manifest, and protections work, regardless of intent.

Though the implications of public health and cybersecurity data sharing are different in some instances, the individual liberties and private sector interests intruded upon are significant in both contexts. Public health initiatives at the extreme interfere with freedom of movement—for example quarantines—and bodily integrity with forced treatment for recalcitrant patients with highly contagious diseases.<sup>60</sup> Public health information sharing at times divulges intimate health information—including data about sexual practices, sexual partners, and drug use—to health officials, and in some instances others who are at risk of infection. Such sharing intrudes on individual privacy, questions the sanctity of the doctor-patient relationship, reveals intimate associations, and places burdens on private health care providers. Public cybersecurity information sharing activities—depending upon the information being shared—may reveal private communications, associational interests, the physical whereabouts and movements of individuals, and other personal details, and they may also disclose confidential information about companies' networks and policies.

Much of the data under discussion in the cybersecurity information sharing debates is held by the private sector. The sharing of such data may impose direct administrative costs on firms, as well as create risks to their competitiveness by forcing firms to reveal internal practices and strategies, and market reputation. This is true in the field of public health as well. Much of the data that fuels public health initiatives comes from private entities, and some are collected and sold by private organizations (e.g., insurance companies). Although there are public good benefits derived by sharing such data with the government, there are proprietary interests at stake too.

Finally, there are multifaceted, sometimes competing, national security concerns in both domains. Concerns about bioterrorism at times lead the government to limit the sharing of detailed health vulnerabilities, scientific information (i.e., viral structure and information that could allow for artificial replication),<sup>61</sup> or information relating to research, stockpiles, or

---

60. For instance, Directly Observed Therapy (DOT) may be used as a compulsory compliance-enhancing strategy when highly infectious individuals have a history of non-compliance. LAWRENCE GOSTIN, *PUBLIC HEALTH LAW: POWER, DUTY, RESTRAINT* 417 (2d ed. 2008).

61. Denise Grady & William J. Broad, *Seeing Terror Risk, U.S. Asks Journals to Cut Flu Study Facts*, N.Y. TIMES (Dec. 20, 2011), <http://www.nytimes.com/2011/12/21/health/fearing-terrorism-us-asks-journals-to-censor-articles-on-virus.html>.

response preparation activities—although it could be useful for public health purposes.<sup>62</sup> Again, this parallels the cybersecurity environment where the government is both pressing for greater information sharing to improved cybersecurity and seeking an informational advantage to support offensive cyber activities.

Like all analogies, the one between public health and public cybersecurity has limitations. However, we maintain that many of the objections to the analogy are more limited and nuanced than they first appear, and that regardless, the analogy provides important lessons about the potential benefits of cybersecurity information sharing and the conditions and mechanisms for its success. From this foundation we can better consider the role information sharing can play in supporting public cybersecurity goals, and better envision the robust protections and governance models necessary to support it in a manner consistent with other values.

#### B. THE ROLE OF INFORMATION SHARING IN PUBLIC HEALTH

Public health is facilitated by a wide range of interventions at the local, state and federal level, many of which are fueled by data. Data informs and makes possible many of the activities necessary to advance public health including: (1) preventing the spread of diseases, epidemics, injuries, and protecting against environmental hazards; (2) promoting healthy behaviors; (3) responding to health incidents and environmental hazards, and assisting communities during recovery; and (4) assuring the quality and accessibility of public health services.<sup>63</sup> All are supported by research, which requires data on population level health and disease that informs the activities in each domain.

In considering the role of information sharing in the public health—and later cybersecurity—context, it is useful as seen in Table 1 to align information sharing with particular activities under two large strategic goals: (1) prevention, which includes promoting healthy behaviors, and (2) management and response.

---

62. There is also a long history of nation states developing and using biological and chemical agents offensively, which adds to potential national security concerns about information sharing from the government's perspective. See W. Seth Carus, *The History of Biological Weapons Use: What We Know and What We Don't*, 13 HEALTH SECURITY 219, 239 (2015).

63. See *The Public Health System and the 10 Essential Public Health Services*, CDC, <http://www.cdc.gov/nphsp/essentialservices.html> (last updated May 29, 2014).

Reviewing the role of information sharing in meeting public health goals helps clarify the potential roles information sharing could play in advancing public cybersecurity.

Table 1: Goals and Associated Activities and How They Relate to Information Sharing

	Preventative Orientation (Reducing Vulnerabilities)	Response Orientation (Managing Insecurity)
Essential public health/public cybersecurity activities (Essential public good production activities)	<ul style="list-style-type: none"> <li>• Improving artifacts</li> <li>• Community and individual empowerment</li> <li>• Policy development</li> </ul>	<ul style="list-style-type: none"> <li>• Detection</li> <li>• Identification</li> <li>• Containment</li> <li>• Treatment</li> </ul>
← Ongoing Research Activities →		
Role of Information	<ul style="list-style-type: none"> <li>• Program evaluations for efficacy</li> <li>• Inform changes to laws, regulation, architecture, and programs</li> <li>• Educate public</li> </ul>	<ul style="list-style-type: none"> <li>• Disease or symptom surveillance</li> <li>• Investigate outbreaks</li> <li>• Contact tracing and spread of disease</li> <li>• Decisions regarding future preventative activities</li> </ul>

### 1. Prevention

In public health, vaccine programs, institutionalizing sanitation infrastructure, occupational hazard laws, behavioral regulations such as seatbelt laws, education, and behavioral incentive programs reduce susceptibility to disease and injury. Information is collected and shared to examine scientific and cultural or structural artifacts that introduce public health vulnerabilities within the population and particular communities. Preventative efforts include public education, community empowerment, artifact improvement, and policy development such as the creation of vaccine programs. All of these activities benefit from basic research to understand causes and effective methods<sup>64</sup> and are further informed by project evaluation, both of which require data collection and sharing. The ability to share data between distributed public health actors enables and sustains coordinated actions—for example, allowing public education initiatives to focus on particularly at-risk populations, or widely disseminating particularly effective interventions.

---

64. *Id.*

## 2. *Management and Response*

Public health activities also identify and manage the constantly changing landscape of disease. Doing so requires monitoring occurrences of disease, providing information to healthcare practitioners and individuals, and, where possible, eradicating root causes of disease. Disease eradication is almost never achieved,<sup>65</sup> so even mitigated health threats may return. Viruses like influenza evolve over time, and rare avian and swine strains may cross over and become capable of human-to-human transmission, which introduces new threats that must be detected and identified among other common influenza strains.<sup>66</sup>

Management and response are information intensive public health activities. While nearly all public health activities benefit from data that can inform interventions and assist in program evaluation, disease detection presents particularly intense information demands. Whether it is monitoring the level and spread of well-known infections—such as HIV—or identifying early signs of a new virus, public health relies on a massive and distributed disease surveillance infrastructure. National data surveillance systems are an important component of the public health information sharing ecosystem, especially within management and response. Centers for Disease Control and Prevention (CDC) officials define public health surveillance as the “systematic, ongoing collection, management, analysis, and interpretation of data followed by the dissemination of these data to public health programs to stimulate public health action.”<sup>67</sup> The purposes of health surveillance systems are made clear to justify them and distinguish them from other data collecting activities. For example, the CDC National Center for Chronic Disease Prevention and Health Promotion states that it engages in surveillance activities to: (1) understand risk behaviors, preventive care practices, and the burden of selected chronic diseases; (2) monitor the progress of current prevention efforts; and (3) inform policy and public health decisions.<sup>68</sup>

Health data surveillance assists in detecting vulnerabilities and vectors of disease—for example, the source of a food-borne illness—and new

---

65. See Richard J. Whitely, *Smallpox: A Potential Agent of Bioterrorism*, 57 ANTIVIRAL RES. 7, 8 (2003).

66. *H5 Viruses in the United States*, CDC, <http://www.cdc.gov/flu/avianflu/h5/index.htm> (last updated Aug. 5, 2015).

67. Stephen B. Thacker, Judith R. Qualters, & Lisa M. Lee, *Public Health in the United States: Evolution and Challenges*, 61 CDC MORBIDITY & MORTALITY WKLY. REP. (SUPP.) 3, 3 (July 27, 2012), <http://www.cdc.gov/mmwr/pdf/other/su6103.pdf>.

68. *Chronic Disease Prevention and Health Promotion: Statistics and Tracking*, CDC, <http://www.cdc.gov/chronicdisease/stats> (last visited May 20, 2015).

threats—such as novel virus strains—but the data collection necessary to support these activities can be quite detailed and invasive. Mechanisms and policies to balance values and tradeoffs are of the utmost importance; otherwise individuals, or the health care providers who serve them, may take evasive measures to limit the collection of sensitive information. Such evasive measures would jeopardize the overall benefits derived from this surveillance. The data garnered by public health surveillance systems are foundational to essential public health activities. A response orientation depends upon event detection and threat identification. These activities directly benefit from coordinated data surveillance activities. Data surveillance has successfully identified otherwise invisible threats like the spoiled vaccine source causing poliomyelitis in 1955 and novel infections like SARS corona virus and variants of influenza.<sup>69</sup> Though we consider the activities enabled by public health surveillance to fall under management and response, information generated by these ongoing activities feeds future preventative efforts and programs. Data about program effectiveness or changes within the security ecosystem fuel better prevention mechanisms through improved education efforts, program improvement, and the formation of more targeted policies.<sup>70</sup>

Modern public health initiatives rely heavily upon data generated by active and passive<sup>71</sup> public health surveillance systems,<sup>72</sup> which depend on widespread and systematic information sharing. Data surveillance within public health occurs at many levels with varying degrees of specificity. The collection of public health and disease surveillance information in the United States is largely conducted at a state level, but data are often shared at the national level to facilitate consideration across the larger population.<sup>73</sup> Surveillance may be of specific chronic conditions or

---

69. GOSTIN, *supra* note 60, at 290.

70. *The Public Health System and the 10 Essential Public Health Services*, *supra* note 63.

71. Passive surveillance is characterized as information reported to public health agencies by healthcare providers or laboratories, whereas active surveillance involves the solicitation of information by public health officials from healthcare providers or laboratories. Contact tracing (identifying others an infected individual may have contracted an illness from or spread the illness to) is a common active surveillance practice. RUTH GAARE BERNHEIM ET AL., *ESSENTIALS OF PUBLIC HEALTH ETHICS* 98 (2013).

72. There is a famous quote by U.S. Surgeon General Dr. David Satcher (1998–2002) in which he says “In public health, we can’t do anything without surveillance. That’s where public health begins.” *Id.* at 99.

73. There are notable exceptions, including many federally administered surveys. However, many national databases originate from state-run, cooperative programs. *See Summary of NCHS Surveys and Data Collection Systems*, CDC, <http://www.cdc.gov/nchs/>

infectious diseases, or may be at a more general level of capture like morbidity and birth data (also known as vital statistics). Surveillance may also be behavior specific. The CDC administers the Behavioral Risk Factor Surveillance System (BRFSS), which operates a voluntary telephone survey examining health-related risk behaviors, chronic health conditions, and use of preventative services.<sup>74</sup>

Data surveillance systems aim to produce consistent data over time, thus enabling historical study and comparison. These systems are uniquely situated to provide vital information for preventative efforts and programmatic decision-making, but they also fuel most of the response-oriented activities (detection, identification, containment, and treatment). The systematic collection of morbidity data across states is attributed to reducing the window of identifying incidents of natural and man-made disease. For example, the morbidity and case clusters of unusual pneumonia and rare cancers in 1981 led to the discovery of AIDS.<sup>75</sup>

The information shared to support these public health activities varies. Some collected and shared information presents little risk to other values. For example, the CDC coordinates PulseNet,<sup>76</sup> a network of state and local laboratories that analyzes the DNA fingerprints of bacteria that cause gastrointestinal infection, often associated with food-borne illness. PulseNet helps coordinate outbreak detection by creating a tailored information sharing program with minimal impact on privacy or other values. When a patient seeks medical care for severe food poisoning, healthcare providers take a fecal sample and send it to a network lab for analysis. By profiling the DNA of the underlying cause of infection (bacteria) and registering these data on a shared database, epidemiologists are able to track outbreaks or identify the source of food contamination and possibly initiate a product recall.<sup>77</sup> When a serious epidemic is detected, the CDC works with local and state public health officials to stop the spread and make announcements to the public about the incidents. This was recently demonstrated when drug-resistant *Shigella*

---

data/factsheets/factsheet\_summary.htm (last updated Mar. 6, 2015). The organization and coordination between these stakeholders are discussed below. *See infra* Part IV.B.

74. *Behavioral Risk Factor Surveillance System*, CDC, <http://www.cdc.gov/brfss/> (last updated Sept. 15, 2015).

75. GOSTIN, *supra* note 69, at 292.

76. *PulseNet*, CDC, <http://www.cdc.gov/pulsenet> (last updated Sept. 9, 2013).

77. Only when epidemiologists and microbiologists detect unusual patterns do public health officials seek out more information related to an outbreak using personal interviews that potential sources of contaminated food. *See Frequently Asked Questions, PulseNet*, CDC, <http://www.cdc.gov/pulsenet/about/faq.html> (last updated July 22, 2013).

*sonnei* began spreading within the United States and was announced in the Morbidity and Mortality Weekly Report (MMWR) with synthesized information for the public and public health practitioners.<sup>78</sup>

These measures to detect, identify, and respond to outbreaks work because the CDC plays a coordinating role in the analysis of samples, in information sharing between public health officials and laboratories, and in response formulation. Responsibilities for detecting these infections are distributed among many stakeholders, but are made possible on a national scale due to federal coordination.

There are other programs that demand and share far more sensitive information, intruding more heavily on individual privacy and posing a greater risk to data subjects if information is misused. We will explore how these systems are designed in the following section.

#### IV. PROMOTING GOALS AND MEDIATING AMONG VALUES: INSIGHTS FROM PUBLIC HEALTH

Data sharing occurs in the context of complex commitments to other values—particularly patient privacy and maximal participation in the health care system—that are at times in tension with public health information needs. Information sharing activities, if not carefully constructed, risk undermining the accuracy and completeness of the datasets since fear of stigma or discrimination stemming from keeping identified records can discourage individuals from seeking care and thus being recorded in the first place.<sup>79</sup> These datasets are often crucial to tailoring and evaluating interventions, so incomplete or misleading datasets have important public health consequences. A web of ethical guidelines, laws, policies and practices mitigates these tensions.

Public health information sharing arrangements are guided by a set of ethical principles. First is a commitment to seek the information necessary to implement effective policies and programs. Second is a commitment to provide communities with information necessary to make decisions on policies or programs and to facilitate community participation and consent

---

78. Anna Bowen et al., *Importation and Domestic Transmission of Shigella sonnei Resistant to Ciprofloxacin—United States, May 2014–February 2015*, 64 CDC MORBIDITY & MORTALITY WKLY. REP. 318, 318–20 (2015), <http://www.cdc.gov/mmwr/pdf/wk/mm6412.pdf>.

79. As one example, fear of social stigma and discrimination resulting from reporting requirements has been shown to keep HIV positive individuals from initially getting tested and seeking early treatment. Margaret A. Chesney & Ashley W. Smith, *Critical Delays in HIV Testing and Care*, 42 AM. BEHAV. SCI. 1162, 1162 (1999), <http://abs.sagepub.com/content/42/7/1162.full.pdf+html>.

in program implementation. Third is a commitment to make information held by public health institutions available in a timely manner consistent with relevant mandates and resource constraints. Fourth is a commitment to protect the confidentiality of information that can bring harm to an individual or community, and to limit intrusions on confidentiality to instances where there is high likelihood of significant harm to the individual or others. Data sharing procedures are informed by the additional principles of accountability, stewardship, scientific practice, efficiency, and equity.<sup>80</sup>

Though these key principles of public health information sharing aim to support common goals, in practice they can be in tension. Seeking and making information accessible to facilitate community decision making can erode individual privacy and harm individuals or communities, for example where a community suffers economic losses due to fear of a contagious disease known to be affecting the community.<sup>81</sup> Systematic data collection through disease surveillance places these principles in tension too. Public health surveillance systems, in general, do not rely on patient consent to collect incident data but nonetheless take great care to protect individual privacy through policies, practices, and technical mechanisms. This practice reflects a policy decision that citizens have a social contract and duty to inform the rest of the state of where their health directly implicates the well-being of others.<sup>82</sup> In this context, relying on voluntary participation or even offering an “opt-out” would undermine the health of society as a whole, so privacy loss is tolerated, but mitigated.

These broad ethical guidelines shape the information sharing policies, practices and mechanisms across public health institutions. Their impact is evident in legal provisions, contractual agreements, the construction of standards and data sets, and the design of platforms that facilitate access to public health data. Below we review legal frameworks, institutional policies, and mechanisms that shape the public health information sharing environment.

---

80. See CDC, CDC-GA-2005-14, CDC/ATSDR POLICY ON RELEASING AND SHARING DATA 5–6 (2005), <http://www.cdc.gov/maso/Policy/ReleasingData.pdf>.

81. Fred Barbash, *Ebola-Stricken Liberia is Descending into Economic Hell*, WASH. POST (Sept. 30, 2014), <http://www.washingtonpost.com/news/morning-mix/wp/2014/09/30/hit-by-ebola-liberia-is-descending-into-economic-hell>.

82. See Lisa M. Lee, Charles M. Heilig, & Angela White, *Ethical Justification for Conducting Public Health Surveillance Without Patient Consent*, 102 AM. J. PUB. HEALTH 38, 41 (2012).

We organize our review around a set of four overarching principles that we distilled from the information sharing policies and practices that have emerged under the ethical guidelines outlined above: (A) expert and collaborative decisions about data collection and governance, (B) reporting minimization and decentralization, (C) earliest de-identification, and (D) limitations on non-public health related uses, particularly limits on public record requests and law enforcement access and use. These principles facilitate effective large-scale public health information sharing activities, while protecting other values that might, if left unchecked, undermine public support for public health driven data collection, and suppress access to care, thus reducing the availability of essential data. Lastly, we review the access policies and mechanisms that support public health information sharing to explore how they influence, practice, and further promote the four core ethical principles.

Together, the principles and aligned access models provide useful guidance for cybersecurity policy and practice. Tailoring information sharing to support public cybersecurity goals, and implementing policies and mechanisms aligned with these principles, could assuage the concerns of individuals and organizations who might otherwise attempt to subvert cybersecurity data collection and sharing.

A. EXPERT AND COLLABORATIVE DECISIONS ABOUT DATA  
COLLECTION AND GOVERNANCE: PRACTICES, STANDARDS, AND  
RELEASE PROCEDURES

The capacity for public health databases and data collection and sharing mechanisms to advance public health goals depends upon their utility and interoperability, and demands coordination and data governance at a national level. However, these decisions cannot be made by federal entities alone. Public health experts develop decisions about information sharing initiatives, data elements, practices, technical standards, and the associated financial and technical burdens, along with organizational responsibilities. These decisions evolve over time to respond to emerging needs and feedback from practitioners, public officials, and the general public.

Choices about what data should be collected and at what granularity impact future public health utility and present both administrative and privacy tradeoffs. The Council of State and Territorial Epidemiologists (CSTE)<sup>83</sup> works directly with the CDC to determine which diseases

---

83. COUNCIL OF STATE & TERRITORIAL EPIDEMIOLOGISTS, <http://www.cste.org> (last visited July 20, 2015).

should be included or removed from national reporting, like those included in the National Notifiable Disease Surveillance System. For federally administered surveillance surveys (e.g., BRFSS), states and public health partners are able to request new data elements or topic specific modules in order to improve the survey utility for stakeholders' public health activities.<sup>84</sup> Stakeholder feedback on nationally administered activities builds trust and cooperation among public health partners, and improves the utility of the survey data for research and program/initiative evaluations. Sub-committees, made up of program officers from the CDC as well as CSTE, have been convened to perform surveillance oversight and evaluation,<sup>85</sup> as well as create national plans for data governance like the CDC-CSTE Intergovernmental Data Release Guidelines.

At the federal level, the CDC/ATSDR Policy on Releasing and Sharing Data governs data quality, compliance, and data release and sharing.<sup>86</sup> The policy tasks the Chief Information Officer (CIO) with evaluating data quality and the risk of disclosing private or confidential information, and establishing obligations for non-CDC data users, grantees, contractors, and partners, among other things.<sup>87</sup> When assessing data quality, CIOs are required to test for completeness, validity, reliability, and reproducibility.<sup>88</sup> The CDC follows quality guidelines put forth by itself, Health and Human Services (HHS), and the Office of Management and Budget (OMB).<sup>89</sup> The HHS Information Quality Guidelines stipulate that Requests for Correction (RFC) and Requests for Reconsideration (RFR) may be submitted to HHS for review, and

---

84. Amy B. Bernstein & Marie Haring Sweeney, *Public Health Surveillance Data: Legal, Policy, Ethical, Regulatory, and Practical Issues*, 61 CDC MORBIDITY & MORTALITY WKLY. REP. (SUPP.) 30, 33 (July 27, 2012), <http://www.cdc.gov/mmwr/pdf/other/su6103.pdf>.

85. See Lisa M. Lee & Stephen B. Thacker, *The Cornerstone of Public Health Practice: Public Health Surveillance, 1961–2011*, 60 CDC MORBIDITY & MORTALITY WKLY. REP. (SUPP.) 15, 16 (Oct. 7, 2011), <http://www.cdc.gov/mmwr/pdf/other/su6004.pdf>.

86. See CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80.

87. The CIO must report the implementation to the CDC Associate Director for Science (ADS) as well. *Id.* at 10.

88. *Id.* This requirement extends to research publications, official reports, oral presentations, and statistical information (i.e. data) put out by the CDC, but does not apply to documents authored or presented by other non-CDC parties. *Id.* at 3–4.

89. See *Advancing Excellence & Integrity of CDC Science*, CDC, <http://www.cdc.gov/od/science/quality/support/info-qual.htm> (last visited May 1, 2015); OMB Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies, 67 Fed. Reg. 8452 (Feb. 22, 2002), <https://www.whitehouse.gov/sites/default/files/omb/fedreg/reproducible2.pdf>. OMB often mandates the use of specific questions for surveyed variables, like sex, ethnicity, and race. Bernstein & Sweeney, *supra* note 84, at 33.

requires that these requests are posted with all documentation and status updates on the Internet for public transparency.<sup>90</sup> The mechanism provides recourse for those concerned with the quality of data released for public use. Requestors range from contract employees to advocacy or trade organizations and private citizens.

CDC guidelines require that data stewards review all data prior to release to assess the risks of re-identification and determine if additional steps are necessary to ensure confidentiality. This evaluation of risk points to the 18 variables considered identifiers under the Health Insurance Portability and Accountability Act (HIPAA) that must be removed before a dataset may be considered de-identified<sup>91</sup>—even though the policy guidance notes that releasing public health information is not covered under HIPAA.<sup>92</sup> The policy notes the tension between reducing the privacy risk of disclosure, and managing the overall utility of the data for public health research and practice. The U.S. Census Bureau provides additional resources covering Statistical Disclosure Control that other agencies may adapt to minimize risks when releasing data.<sup>93</sup> Occasionally the CDC is unable to specify formats, delivery modes, and opportunities for data sharing and release. Pre-existing funding and cooperation agreements for surveillance activities can reduce their ability to influence data products and uses. In contrast, when a contract dictates funding, it is easier to influence and evolve the data specifications and sharing obligations, including privacy requirements, in a way that benefits public health.<sup>94</sup> The centralized authority from the CDC and other coordinating groups, based on stakeholder feedback, provides guidance and contract incentives to make data as open and accessible as possible while balancing privacy and sensitivity considerations.

Decisions about the release of sensitive information do not only occur at a national level. In addition to the practices described above, the National Center for Health Statistics (NCHS) (a federal entity) and the

---

90. *Information Requests for Corrections and HHS' Responses*, DEP'T HEALTH & HUMAN SERVS., <http://aspe.hhs.gov/information-requests-corrections-and-hhs-responses> (last visited Oct. 20, 2015).

91. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 11.

92. CDC, CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES FOR RE-RELEASE OF STATE-PROVIDED DATA 71 (2005), <http://stacks.cdc.gov/view/cdc/7563>.

93. *Statistical Disclosure Control (SDC): Documents used by the Census Bureau's Disclosure Review Board*, U.S. CENSUS BUREAU, <http://www.census.gov/srd/sdc/> (last visited May 1, 2015); *see also* U.S. CENSUS BUREAU, DISCLOSURE REVIEW BOARD (2001), <http://www.census.gov/srd/sdc/wendy.drb.faq.pdf>.

94. Bernstein & Sweeney, *supra* note 84, at 34.

National Association for Public Health Statistics and Information Systems (a non-profit that represents states and territories) collaboratively review researchers' data requests for restricted vital statistics files. The National Association for Public Health Statistics and Information Systems reviews the requests before the federal entity, NCHS, which allows the state data owners to share oversight with the federal government.<sup>95</sup> The distribution of responsibility and oversight adds an additional layer of protection and collaboration between public health stakeholders.

The CDC and the Agency for Toxic Substances and Disease Registry (ATSDR)<sup>96</sup> work with other public coordinating groups and periodically amend current practices and release guidance documents on data use, release, and sharing. These guidance documents clarify goals in data management and sharing practices, ensure compliance with relevant federal laws and guidelines (e.g., HIPAA, the Freedom of Information Act (FOIA),<sup>97</sup> OMB Budget Circular A110, and Information Quality Guidelines, etc.), and promote the routine and prompt sharing of data by the federal government with public health partners while protecting sensitive data. The data covered by federal guidance documents does not include data owned by private organizations and shared with the federal government, though these data may still fall under the jurisdiction of other laws, regulations, or agreements.<sup>98</sup>

Technical standards and requirements facilitate information sharing, and the protection of privacy and other values. Public health policy looks to develop voluntary consensus standards to facilitate information sharing.<sup>99</sup> The Public Health Information Network (PHIN) is a national initiative within the CDC Division of Health Informatics and Surveillance (DHIS) designed to increase the capacity of public health agencies to electronically exchange data and information through the establishment of standards<sup>100</sup> and technical requirements.<sup>101</sup> Most of the

---

95. *Id.*

96. The CDC and ATSDR are both under the HHS. Many of the data sharing policies were written jointly by both of these agencies since both agencies play a large role in public health data collection and dissemination. For simplicity in this paper, we will refer only to the CDC when talking about public health data sharing practices.

97. For a discussion of FOIA, see *infra* Part IV.D.

98. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 3.

99. To develop voluntary consensus standards, the federal government fulfills requirements set forth by the National Technology Transfer Advancement Act of 1995 (NTTAA). See generally National Technology Transfer Advancement Act of 1995, Pub. L. No. 104-113.

100. PHIN uses OMB Circular A-119 for their definition of "standard." *Standards and Interoperability Enterprise Services*, CDC, <http://www.cdc.gov/phinf/resources/>

standards for public health data are directed by existing laws and policies that specify the voluntary consensus and evaluation processes, and are enumerated on the PHIN website.<sup>102</sup> These PHIN standards and interoperability activities are part of CDC-wide standardization activities, which the National Institute of Standards and Technology (NIST) publishes annually in the NTTAA annual reports.<sup>103</sup>

The CDC has a goal to make data standards and documentation compatible with those used in private industry to facilitate data use for public health purposes. Given the often rapid pace of innovation within the private sector, these standards are developed and reviewed for best practices, and the CDC recommends data documentation elements in its data sharing policy.<sup>104</sup>

Through collaboration across industry and government, public health officials have designed interoperable data formats, systems, and policies that improve the potential utility of information sharing activities undertaken to promote public health goals. These collaborative efforts foster trust across institutional actors and the public, and support innovation within the field. Officials are also able to attend to values such as privacy through policy measures and technical choices that affect the entire ecosystem.

#### B. REPORTING MINIMIZATION AND DECENTRALIZATION

Reporting minimization and decentralization are common elements of the public health data collection landscape. Legal frameworks, institutional policies and practices, and technical approaches to data sharing reflect preferences for keeping data in the hands of the initial collector rather than pooling it, and minimizing the data that flows when sharing is necessary. Adherence to these principles erects practical barriers to the misuse or repurposing of public health data at scale; multiple

---

standards/index.html (last updated July 1, 2015); *see also* OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB CIRCULAR NO. A-119, REVISED, FEDERAL PARTICIPATION IN THE DEVELOPMENT AND USE OF VOLUNTARY CONSENSUS STANDARDS AND IN CONFORMITY ASSESSMENT ACTIVITIES (1998), <http://www.nist.gov/standardsgov/omb119.cfm>.

101. Public Health Information Network Homepage, CDC, <http://www.cdc.gov/phn/about/index.html> (last updated Sept. 10, 2015).

102. *Data Interchange Standards*, CDC, [http://www.cdc.gov/phn/resources/standards/data\\_interchange.html](http://www.cdc.gov/phn/resources/standards/data_interchange.html) (last visited May 1, 2015).

103. Database of Reports Submitted Under the National Technology and Advancement Act of 1995, NIST, <https://standards.gov/NTTAA/agency/index.cfm?fuseaction=agencyReports.main> (last updated Mar. 7, 2013).

104. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 3.

systems must be compromised, or multiple entities convinced for a shift in use to occur. When breaches or shifts in use occur, the limited nature of the data often reduces the potential harms. Minimization can reduce the attractiveness for abuse of the underlying data by limiting its potential for misuse or repurposing.

Much of the data used for public health purposes is not collected or held at the federal level, but rather generated, stored, and used by state, local or non-state organizations. Data obtained for public health uses come from four different types of sources: (1) data that the CDC collects directly using federal funds, (2) data that other agencies or organizations collect for the CDC (e.g., through procurement mechanisms like grants, contracts or cooperative agreements),<sup>105</sup> (3) data that other organizations like state health departments report to the CDC, and (4) privately collected data shared with the CDC. As discussed below, data that parties collect under federal or state authority to advance specific public health goals may only be used for these purposes.<sup>106</sup>

The Public Health Services Act (PHSA) authorizes federal public health data collection.<sup>107</sup> The government often uses federal administrative data, including Medicare, Medicaid, and Social Security Disability data, for public health surveillance purposes.<sup>108</sup> Many data reporting mechanisms are voluntary collaborations between data holders (often state health departments) and the federal government. While most state data reporting to the federal government is voluntary, it is conducted with federal and peer-based committee guidance via the CSTE on data collection, standardization, and compliance with state laws and regulations.<sup>109</sup>

When data is collected at a federal level, only data necessary to achieve public health goals are reported. Federal agencies like the CDC have their own collected datasets (like survey responses) that may include identifiable information, but these datasets are limited. There are notable emergency cases where the federal government requires identifiable data, such as in bioterrorism responses that require joint law enforcement and public health action using special information sharing protocols that comply with

---

105. *Id.*

106. Data may be used for other purposes only if the data subject gave appropriate consent at the time of collection. *See infra* Part IV.D.

107. 42 U.S.C. §§ 242b, 242k, 242l (2012).

108. Bernstein & Sweeney, *supra* note 84, at 32.

109. GOSTIN, *supra* note 60, at 286–87.

all applicable laws and regulations.<sup>110</sup> In most cases, identifiable public health surveillance data are only maintained at the local government level (i.e., state or county) where it was obtained. Local and state laws regulate collection and confidentiality as well. These state and local entities are ultimately responsible for ensuring confidentiality protections to the data they collect and maintain.<sup>111</sup> This separation between collection and reporting means that most sensitive micro-data never reach the federal level, which is where the majority of data releases and sharing activities in support of public health occur.<sup>112</sup>

This separation between collection and federal reporting, as well as clear delineation about what micro-data are appropriate for public health uses, is vital to making these national reporting structures work while balancing the rights of individuals and the benefits for the collective. As a result, public health data is often reported in a *relatively* privacy-protective manner. Often, no identifiers and only broad regional locations are reported as summary level statistics. Though there are perennial concerns that someone can easily re-identify data when coupling with other attributes (e.g., age, geolocation, etc.), efforts to remove identifiable data, along with limited federal collection, assists in protecting the privacy of citizens. This network of information providers and targeted federal collection activities make possible the robust data available for public health activities.

### C. EARLIEST FEASIBLE DE-IDENTIFICATION

At times advancing public health goals requires sharing identifiable information that allows officials to link these data to other datasets or identify persons with a specific disease or health condition. In almost all cases, these identifiable data only remain at the level where the

---

110. Joint investigations between law enforcement and public health officials imply that both entities may be interviewing (and obtaining data about) potential patients, and that public health officials may need to disclose protected health information to law enforcement to avert a serious threat to health or safety as guided under 45 C.F.R. § 164.512(j) (2013). Many emergency information sharing protocols are issued at a local level to ensure that all applicable laws and regulations (including state) are complied with in a timely and orderly fashion. *See, e.g.*, L.A. CNTY. DEPT OF HEALTH SERVICES, L.A. CNTY. SHERIFF'S DEPT & FBI L.A. FIELD OFFICE, JOINT BIOTERRORISM INVESTIGATION MEMORANDUM OF UNDERSTANDING (2005), <http://www2a.cdc.gov/PHLP/docs/joint%20mouLA.pdf>.

111. CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 6; *see also* CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 8.

112. Bernstein & Sweeney, *supra* note 84, at 30.

intervention occurred, which is usually the state or local level.<sup>113</sup> In limited cases, such as a rare disease outbreak or certain high-risk disease surveillance programs, these local or state entities may share identifiable data with other jurisdictions or report them to federal agencies. For example, within the HIV/AIDS surveillance system, experts support the routine sharing of some data with identifiers in order to resolve duplicate case counts across states and territories to assure data quality at a national level.<sup>114</sup>

In the cases where identifiable data must be transferred, there are policies in place to limit risk. Encryption standards and practices—such as replacing names with numbers in records and maintaining the file that connects them separately and in an encrypted format—aim to reduce the potential risk these sharing mechanisms impose.<sup>115</sup> However, in most cases where an organization collects identifiable data, it is de-identified as soon as possible, and before sharing occurs.

The commitment to earliest feasible de-identification plays a particularly important role in public health reporting obligations. Obtaining patient consent to share data for public health reporting would add an administrative burden to healthcare professionals, potentially slow down an already cumbersome reporting process (timeliness is particularly prized in some settings, such as when a new communicable disease is spreading), and limit reported data. Where compulsory information sharing is necessary for public health purposes, de-identification and other efforts at minimization (described above) are largely accepted as sufficient to mitigate privacy harms.<sup>116</sup>

Balancing potential utility with individual privacy is an ongoing struggle as data reporting needs and systems evolve. Though much of the data that is reported at a national level is de-identified to some extent, datasets have varying levels of specificity, and some may be tied to additional data that makes identification easier, such as a geographic marker of residence.<sup>117</sup> Geographic markers, gender, age and other data

---

113. *Id.*

114. Amy L. Fairchild et al., *Public Goods, Private Data: HIV and the History, Ethics, and Uses of Identifiable Public Health Information*, 122 PUB. HEALTH REPS. (SUPP.) 7 (2007), <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1804110/pdf/phr122S10007.pdf>.

115. *Standards to Facilitate Data Sharing and Use of Surveillance Data for Public Health Action*, CDC, <http://www.cdc.gov/nchhstp/programintegration/SC-Standards.htm> (last updated Mar. 11, 2014).

116. See BERNHEIM ET AL., *supra* note 71.

117. COMMITTEE TO REVIEW DATA SYSTEMS FOR MONITORING HIV CARE BOARD ON POPULATION HEALTH AND PUBLIC HEALTH PRACTICE, MONITORING

increase the risk of re-identification.<sup>118</sup> However they may be important to understanding public health risks, assessing the efficacy of interventions, and understanding the limits of collected data. There is a notable tension between the need to protect individuals against re-identification and the need to provide public health officials, researchers, and healthcare providers with enough specificity to act or test correlative hypotheses and enough information to understand the strength and limits of their findings.<sup>119</sup>

#### D. LIMITATIONS ON NON-PUBLIC HEALTH USES THAT NEGATIVELY AFFECT INDIVIDUAL INTERESTS

Public health law provides confidentiality protections that limit the reuse of and access to data collected for public health purposes. These use and access restrictions make the intrusions on individual privacy necessary to advance collective public health goals more palatable. Institutional policies, contracts, and technical mechanisms further limit non-public health uses, particularly those detrimental to individual data subjects.

Data held by the CDC—the primary federal public health agency—is subject to the general federal laws and regulations that govern retention,

---

HIV CARE IN THE UNITED STATES: INDICATORS AND DATA SYSTEMS 14 (Morgan A. Ford & Carol M. Spicer eds., 2012). Within the National HIV Surveillance System, various data elements are captured through proxy indicators that are used to improve longitudinal data and make the system more robust.

118. See Latanya Sweeney, *Simple Demographics Often Identify People Uniquely 2* (Carnegie Mellon Univ., Data Privacy Working Paper No. 3, 2000); see also Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE SYMP. ON SECURITY & PRIVACY 111, 111.

119. Differential privacy, which allows researchers to receive statistically meaningful answers to queries while limiting their ability to determine whether a given individual is in or out of the data set, provides a mathematically rigorous way to specify the trade-off between privacy and utility. iDASH (Integrating Data for Analysis, Anonymization, and Sharing), which is funded by the National Institutes of Health, is developing a statistical health information release toolkit with differential privacy. *SHARE: Statistical Health Information Release with Differential Privacy*, iDASH, <https://idash.ucsd.edu/share> -statistical-health-information-release-differential-privacy (last visited Oct. 21, 2015). The Census Bureau is also using differential privacy. See Erica Klarreich, *Privacy by the Numbers: A New Approach to Safeguarding Data*, QUANTA MAG. (Dec. 10, 2012), <https://www.quantamagazine.org/20121210-privacy-by-the-numbers-a-new-approach-to-safeguarding-data>. The CDC states that those assessing risks associated with public health data release should recommend statistical methods to protect confidential information from being disclosed, such as “suppression, random perturbations, recoding, top- or bottom-coding.” CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 11.

access, and disclosure of personally identifiable information.<sup>120</sup> The CDC complies with the Federal Records Act, which governs the retention, destruction, and archiving of federal records,<sup>121</sup> and it sets additional rules regarding retention of data collected for public health purposes.<sup>122</sup> CDC policies leave local (state and municipal) data retention and destruction requirements (which may be more restrictive than federal standards) up to local agencies to ensure their own compliance after reporting to the federal government.<sup>123</sup> FOIA promotes government transparency and accountability to citizens by allowing individuals to request the release of agency records, and it contains nine exemptions,<sup>124</sup> two of which provide specific protection against the release of sensitive health information.<sup>125</sup> While FOIA serves an important purpose, the exceptions balance government transparency and accountability with public health goals and the privacy protections required to achieve them. The result is a policy framework that protects CDC data tied to an individual (e.g., health behavior survey response) and other sensitive datasets from FOIA release.<sup>126</sup> The Privacy Act of 1974 provides additional protections, preventing the disclosure of personally identifiable information contained

---

120. CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 69.

121. 44 U.S.C. ch. 33 (2012); 36 C.F.R. ch. 12, subch. B (2009).

122. *See, e.g.*, CDC Notice, Republication of Systems of Records, 51 Fed. Reg. 42,449, 42,460 (Nov. 24, 1986) (setting rules regarding retrieval of records collected for determining eligibility under the Tuskegee Health Benefit Program).

123. CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 69.

124. Exemptions from FOIA are found under 5 U.S.C. § 552(b) (2012), which protects files related to national security, trade secrets and commercial or financial information from a person that is privileged or confidential, medical files or other similar files where disclosure would constitute a clearly unwarranted invasion of privacy, and information that is prohibited from disclosure by another federal law. The exemptions are aimed to “protect certain equally important rights of privacy with respect to certain information in Government files, such as medical and personnel records.” S. REP. NO. 88-1219, at 8 (1964).

125. 5 U.S.C. § 552(b)(6) limits the application of FOIA to “personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy.” This has been interpreted to protect medical records. *See, e.g.*, *McDonnell v. United States*, 4 F.3d 1227, 1254 (3d Cir. 1993). 5 U.S.C. § 552(b)(3) limits the application of FOIA where records are specifically exempted from disclosure by another statute that leaves no discretion on the issue; or establishes particular criteria for withholding or refers to particular types of matters to be withheld; and if enacted after the date of enactment of the OPEN FOIA Act of 2009, specifically cites to this paragraph.

126. FOIA exemptions include “personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of privacy.” 5 U.S.C. § 552(b)(6).

in a system of records<sup>127</sup> unless the individual to whom the record pertains consents, it is for a “routine use” defined as one “compatible with the purpose” of collection, or another agency requests it and it is relevant to the investigation of a specific violation of law.<sup>128</sup>

The PHSA provides additional privacy protections for public health data that limit the potential for data to be used in ways that negatively affect individuals.<sup>129</sup> The PHSA closes gaps in other federal laws to protect individuals, and, in some cases, organizations, who may be the subject or contributor of information gathered for public health purposes.<sup>130</sup> The CDC offers general confidentiality assurance provisions for both individuals and establishments by prohibiting the use of data collected for any other purpose than the purpose for which it was collected, unless the individual has consented to the alternative use.<sup>131</sup> Further, any information collected during the course of statistical or epidemiological activities may not be published or released in other form if an individual or establishment is supplying the information or described within it are identifiable, unless the individual consents.<sup>132</sup> Confidentiality assurances afforded under the Act protect against disclosure under a court order, and extend protections to institutions and not just individuals. This confidentiality protection allows the CDC to guarantee participants and institutions that their data will only be shared with entities listed on the consent form or Assurance of Confidentiality Statement for the project, which is especially important when data gathering includes sensitive information that otherwise might be withheld, like sexual behaviors, drug

---

127. Unlike FOIA, which pertains to all federal agency records, the Privacy Act pertains only to records within a system of records in which the primary method of data access is through retrieval by full names, social security numbers, or other identifying particulars. 5 U.S.C. § 552a(4).

128. 5 U.S.C. § 552a(b). There are additional limitations—notably the records of deceased persons or non-US citizens are not protected. *Id.*; see also CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 75–77.

129. Public Health Service Act, Pub. L. No. 78-410, 57 Stat. 682 (1944) (codified at 42 U.S.C. ch. 6A); 42 U.S.C. §§ 241(d), 242m(d).

130. Here we focus on the protections offered at a federal level, but it should be noted that since most identifiable data are collected at the local level, state laws are highly relevant. All states offer some legal protection for government-held public health data (especially sensitive data like sexually transmitted infections or data relating to drug and alcohol treatment), but vary in their scope, specificity, and reach to protections for privately held data. It is beyond the scope of this paper to discuss the nuances and shortcomings of these laws, but we acknowledge the weaknesses this introduces into public health data privacy protections at a system level. GOSTIN, *supra* note 60, at 326.

131. 42 U.S.C. § 242m(d).

132. *Id.*

uses, mental health status, or other information that could damage an individuals' reputation financially or socially. These provisions cover research and non-research activities that the CDC carries out or that are under contract to the CDC.

To cover public health information activities conducted under grants or cooperative agreements, the CDC can provide Certificates of Confidentiality to a research project.<sup>133</sup> Certificates of Confidentiality authorize researchers to protect the privacy of individuals so that no federal, state, or local civil, criminal, administrative, legislative, or any other proceedings can compel the release of identifying information unless the individual consents.<sup>134</sup> Certificates cover sensitive information including research pertaining to mental health, and the use and effects of alcohol and other psychoactive drugs.

This suite of additional legal protections ensures that robust public health data collection does not undermine access to health care services and protects institutional interests implicated in data sharing. Confidentiality assurances extend not only to individuals but also institutions that may require protection in order to consider sharing data with the government. The additional Certificates of Confidentiality for activities conducted outside of the federal government allow public health officials to offer important prohibitions on otherwise compulsory uses of data (e.g., for law enforcement activities). Without these protections, researchers and public health officials would lose access to highly sensitive data like statistics related to drug-use and addiction patterns.

#### E. PUBLIC HEALTH INFORMATION SHARING MODELS

A rich and diverse set of information sharing models support public health goals. These models provide examples of the principles in action using specific data sharing mechanisms. Below we discuss three common models of information sharing and access that public health agencies offer. We also discuss non-governmental models for data sharing. A similar range of public sector and private approaches could advance public cybersecurity goals.

##### 1. *Access to Federally Held Data*

The CDC operates on the premise that public health goals and scientific achievement are best promoted by releasing or sharing data in an

---

133. 42 U.S.C. § 241d.

134. CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 77.

open, timely, and responsible way with public health agencies, academic researchers, and other private researchers.<sup>135</sup> Federal public health agencies—namely the CDC in the U.S.—must balance timeliness, data quality, and wide dissemination of data with the need to ensure protection of sensitive information. Sensitive information considerations within public health include protecting the privacy of individuals in the dataset, proprietary interests of data sharing partners, national security interests, and law enforcement activities (including misconduct inquiries and investigations).<sup>136</sup> Balancing these interests in different data sets and contexts requires different approaches.

Three general access models have emerged to support public health data sharing: (1) Open Data—no restrictions and public open access, (2) Limited access with restrictions, and (3) No Access except for internal agency use (not eligible for release or sharing). The choice of access model can depend on legal constraints, ethical guidelines discussed above, and community and public input. Here we will explore how the CDC determines and administers data access, along with the sharing mechanisms, protocols, and laws that preside over data use. These models present opportunities and considerations for handling sensitive data in the cybersecurity context.

a) Open Data: No Restrictions and Public Open Access

According to the CDC/ATSDR Policy on Releasing and Sharing Data, all data the CDC collects or holds that are legally eligible for public release should be publically available within a year of being evaluated for quality and shared with public health partners.<sup>137</sup> When releasing public-use data, the CDC follows procedures to ensure that data are consistent with the standards the PHIN has established.<sup>138</sup>

Each data set must have a specific data release plan to address data sensitivities prior to release. These plans include steps to reduce the risk of confidential information disclosure, procedures to ensure release does not interfere with national security or law enforcement activities, protections for proprietary information, a data quality analysis as required by OMB, instructions on appropriate data use for non-CDC users, timely release schedules, and data formats and standards compliance.<sup>139</sup>

---

135. See CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 1.

136. *Id.* at 2.

137. *Id.* at 7.

138. *Id.*

139. *Id.* at 8.

Data shared without restrictions may be released through the CDC Information Center, and shared through the CDC/ATSDR Scientific Data Repository and the associated data dissemination portal CDC WONDER.<sup>140</sup> The CDC WONDER platform offers an example of how these laws and policies come together in practice. Individuals accessing open data sets through the CDC WONDER platform are shown a brief description of the data, along with applicable use restrictions.<sup>141</sup> In addition to stating these restrictions, the agreement sets lower limits on sample size reporting within a set geographical region in a public dataset (i.e., datasets do not report alone nine or fewer death rates within a sub-geographic region), and makes it clear that any attempt to identify individuals within the dataset is illegal. The platform informs users that they should not further disclose any inadvertent discoveries, and that they must report these discoveries to the NCHS Confidentiality Officer with the contact information provided.<sup>142</sup> Researchers who are found violating the data use restrictions lose access to the CDC platform, and the researcher's institutional sponsors receive notifications about the violation. Access to CDC WONDER is denied until the government conducts an investigation. If the researcher is found to be deliberately making false statements within the jurisdiction of any department or agency of the federal government<sup>143</sup> they may be punished by a fine and/or up to five years in prison.<sup>144</sup>

In some cases, it may be appropriate to release high-level data but offer granular data under a restricted access model. In cases like the National Violent Death Survey, this process has helped improve researcher and program use of these data. Because interested parties can use higher-level data prior to submitting a request for access, they can better determine how data might fit specific needs or research questions.<sup>145</sup>

---

140. *WONDER Online Databases*, CDC, <http://wonder.cdc.gov> (last visited Apr. 1, 2015).

141. *Data Use Restrictions*, CDC, <http://wonder.cdc.gov/DataUse.html> (last visited Oct. 21, 2015). For example, the CDC states that data “may be used only for the purpose for which they were obtained; any effort to determine the identity of any reported cases, or to use the information for any purpose other than for statistical reporting and analysis, is against the law.” *Id.*

142. *Id.*

143. 18 U.S.C. § 1001 (2012).

144. *Id.*

145. *National Violent Death Reporting System: Restricted Access Database (RAD)*, CDC, <http://www.cdc.gov/violenceprevention/nvdrs/rad.html> (last visited May 1, 2015).

## b) Limited Access with Some Restrictions

If data cannot be shared openly with the public, the next policy option is to allow access with restrictions, or to mediate access. Data may be released with restrictions either under controlled conditions or through special-use agreements. Controlled conditions for data release can take the form of research data centers (RDCs) or licenses that limit use of accessed data for non-CDC researchers. Licenses attach these legal responsibilities, binding the external researcher before providing her access to identifiable data.<sup>146</sup> Prior to entering into a special-use agreement, the CDC screens requests to ensure the use is for an appropriate public health purpose.<sup>147</sup>

RDCs offer different access modes, which are not mutually exclusive (a researcher may use a combination of access modes). Researchers interested in using restricted use data must submit a proposal requesting one or more access modes, which include visiting a center<sup>148</sup> or gaining remote access. Approved researchers can remotely query some restricted use datasets held by an RDC. Researchers submit code through an automated system that analyzes the restricted data and returns results. There are technical limitations to these options at the CDC (a researcher can only run some SAS/SAS-callable SUDAAN procedures), for researchers who must use secure email addresses, and for research teams—only one researcher per team can have remote access rights.<sup>149</sup> In addition to time restrictions, research protocols are subject to RDC analysts performing disclosure reviews of the SAS code and output.<sup>150</sup> Any violation or attempt to circumvent remote access protocol to obtain access to prohibited information results in immediate account suspension and potential legal actions. If this method does not meet the researcher's needs, it is possible to apply for staff assisted access, where an RDC analyst runs a set of programs that the researcher created and provides the results separately.

---

146. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 9.

147. *Id.*

148. Researchers may choose between a National Center for Health Statistics (NCHS) RDC (which has three locations in MD, GA, and DC) or Federal Statistical RDCs (which have over 19 locations and are managed by the U.S. Census Bureau). If researchers use a Federal Statistical RDC, a NCHS RDC analyst still handles the proposal and all administrative concerns, and the Census RDC Centers only serve as an access facility. *On Site at a Federal Statistical RDC*, CDC, <http://www.cdc.gov/rdc/b2accessmod/acs220.htm> (last visited May 1, 2015).

149. *Remote Access*, CDC, <http://www.cdc.gov/rdc/b2accessmod/acs230.htm> (last visited May 1, 2015).

150. *Id.*

If allowed, sensitive data may be released under special-use agreements outside the controlled conditions of an RDC. These agreements address co-authorship, but more importantly for this discussion, they provide for the CDC to review all findings resulting from restricted data use, review publications, and establish a time when researchers must return data. In order to be eligible for a special-use agreement, the research must be necessary for a legitimate public health purpose. The agreement must contain a list of data use restrictions, names of all researchers with access to the data, information regarding pertinent laws relating to the use of the data, security procedures and associated penalties for failure to comply, a list of restrictions on releasing data analysis results, procedures for data return to the CDC and managing access of staff changes, and provisions to cover emergency requests for identifiable or confidential data.<sup>151</sup>

These public health procedures and policies, which offer full and access-mediated release of data to promote public good activities, balance data sensitivity concerns with the benefits of open data access. By not restricting access to a binary all-or-none model, public health policy is able to optimize data sharing and use without compromising privacy and security interests.

c) Internal Agency Use Only

One of the core principles within public health information sharing is to make data as accessible as possible with the minimum amount of restrictions necessary to protect individuals and organizations.<sup>152</sup> If the government does not release data, the conclusion is that public use or mediated access modes were not appropriate. Despite the preference for robust access to support public health uses, at times other values council against sharing. Reasons to withhold data may include, but are not limited to: data classified for national security reasons, proprietary data from non-governmental organizations, and identifiable or particularly sensitive data.

The varied role the federal government plays in public health information and data governance, the nuanced options for data access that protect sensitivities while promoting openness and accessibility, and the different approaches to data platform operations, together provide a set of policy and organizational offerings that can be applied to public cybersecurity.

---

151. CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 29–31.

152. *See* CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 80, at 2.

## 2. *Non-Governmental Platforms for Public Health Information Sharing*

In addition to the CDC WONDER platform, there are information sharing platforms managed by non-governmental organizations that operate with limited federal funding and coordination. For example, the BioSense program is a streamlined collaborative data-exchange system that allows users (public health officials in cooperating jurisdictions) access if they agree to contribute funds and share real-time data through a cooperative agreement.<sup>153</sup> It is part of the CDC's National Syndromic Surveillance Program, built in response to a Congressional mandate in the Bioterrorism Preparedness and Response Act of 2002, and later adapted in 2010 to fit broader situational awareness needs of stakeholders.<sup>154</sup>

BioSense aims to provide public health partners with a technology platform to collect and analyze large amounts of health data in a timely manner so that local, state, and federal officials may monitor, detect, and respond to outbreaks and harmful effects from exposure to hazardous conditions.<sup>155</sup> The government distributes funding to public health partners; in 2012 this funding totaled around \$7 million, awarded to the 35 participating health departments.<sup>156</sup> A group of non-governmental organizations runs the system, but the CDC (with input from stakeholders) organized and adapted it for broader use. Users include the CDC, state and local health departments, and other public health partners.

The BioSense 2.0 environment,<sup>157</sup> funded by the CDC, is hosted by the Association of State and Territorial Health Officials (ASTHO). Stakeholder feedback is obtained from the ASTHO in coordination with CSTE, the National Association of County and City Health Officials (NACCHO), and the International Society for Disease Surveillance (ISDS). Other federal agencies including the Department of Defense and

---

153. *BioSense 2.0*, CDC, <http://www.cdc.gov/nssp/biosense/biosense20.html> (last visited Aug. 12, 2015).

154. *BioSense Background*, CDC, <http://www.cdc.gov/nssp/biosense/background.html> (last visited Aug. 12, 2015).

155. *BioSense: Meaningful Use*, CDC, <http://www.cdc.gov/nssp/biosense/meaningfuluse.html> (last visited Aug. 1, 2015).

156. *BioSense: Cooperative Agreement*, CDC, <http://www.cdc.gov/nssp/biosense/cooperativeagreement.html> (last visited Aug. 12, 2015).

157. BioSense 2.0 refers to the latest version of the platform.

Department of Veteran Affairs have assisted in development and integration of the system with existing data systems.<sup>158</sup>

Not all information sharing activities need to be solely funded and administered by the U.S. government. The design of BioSense incentivizes information exchange by making sharing a requirement of participation, and delegating administrative and organizational responsibilities.

## V. RECOMMENDATIONS FOR APPLYING INFORMATION SHARING LAWS AND POLICIES TO CYBERSECURITY

We have identified a set of four principles and three information sharing and access models in the public health field that advance public health goals while mitigating the harm to other individual and collective values. Using the derived principles and access methods, we provide a set of recommendations to guide cybersecurity information sharing. These mechanisms and practices facilitate data access for public cybersecurity activities while balancing the privacy, freedom of expression, innovation, and competitiveness of individuals and organizations.

### A. CLARIFY THE PUBLIC GOALS OF CYBERSECURITY AND THE ROLE OF INFORMATION SHARING IN ADVANCING THEM

The lack of clarity about overarching goals stymies cybersecurity policy generally, and information sharing specifically. Policy makers should adopt the Doctrine of Public Cybersecurity to ensure that information sharing and other initiatives aid in the production of more secure systems and behaviors, and enable management and response of ongoing vulnerabilities. Assuring that our technical infrastructure is able to adequately secure the activities and data we entrust to it is a pressing national priority. Clarifying the aims of national policy would assure that information sharing and other activities are considered for their capacity to advance these dual goals.

### B. CLARIFY CONNECTIONS BETWEEN DATA SHARING PROPOSALS AND PUBLIC GOALS

When advocating information sharing or implementing federal collection, the nexus between the specific information to be shared or collected, its intended use, and relationship to advancing public

---

158. *BioSense: The Community*, CDC, <http://www.cdc.gov/nssp/biosense/community.html> (last visited Aug. 12, 2015).

cybersecurity goals should be clear to contributors and the public. Particularly for data systematically collected or reported through a surveillance system, it is important to make the purpose of collection clear and establish that it will not be used against data subjects for law enforcement or other adverse purposes unrelated to cybersecurity. Uncertainty of end use will negatively impact reporting compliance.

### C. COORDINATE ACTIVITIES USING EXPERT COMMUNITIES

Cybersecurity information needs, including information sharing, require expert guidance and coordination. Public health data surveillance is conducted to advance specific goals under the broad umbrella of prevention, response, and management. While it relies upon both state and federal law and public and private sector actors, the federal government coordinates it. The federal government, with input and feedback from public health partners, facilitates agreement on diseases and problems to target, data to collect and share, controls to protect privacy and other values, and the technical and legal mechanisms to implement these policy decisions. The balanced roles between a wide range of stakeholders, and federal partners, offer a model for public cybersecurity information sharing activities.

While the federal government plays the central role in the public health arena, it is unclear whether that is the appropriate approach for cybersecurity given the distribution of expertise and data. Regardless, the federal government can and should play a role in coordinating data and technical standards. Doing so will promote the overall utility and efficiency of information to achieve public cybersecurity goals, and will ensure that privacy and other interests raised by data sharing are thoroughly and systematically addressed. There is a need for coordination and agreement on standards, what data to report, determination of changing data needs to advance public cybersecurity goals, and management on how to distribute the financial burden of these systems among stakeholders. The success of information sharing relies on input from experts and coordination to reflect differing cybersecurity needs.

Several factors complicate the need for coordination. First, cybersecurity lacks uniform agencies equivalent to health departments at the state level. States have taken different approaches to cybersecurity and distributed responsibilities to different state actors. More importantly, cybersecurity incidents lack clear geographic distinctions, and much useful data is in private, not public, institutions. These factors complicate coordination, sharing, and other information governance responsibilities.

Within the public health context, laws regarding information collection and sharing initially developed on a state-by-state basis, with limited federal coordination. It was difficult at the national level to find coherence among these ad hoc laws. Eventually the federal government took on a stronger coordinating role. Finding enough coherence among these ad hoc laws to implement national information sharing and open data practices took many years.

Aspects of the cybersecurity landscape present additional complications beyond those faced in public health. Advanced coordination, therefore, is particularly important. Expertise is spread across many stakeholders: those who run infrastructure, those who develop tools to defend it, and those who represent the interests of system users. To date, a subset of these experts have driven policy deliberations. In particular, civil society organizations, representing the interests of users and supporting values such as privacy, have been relegated to a largely reactive role. Ensuring that all stakeholders with expertise are able to participate in defining the cybersecurity information sharing ecosystem is key to achieving widespread public support for information sharing in this context. Such inclusion was essential to achieving such public support for information sharing in the public health arena.

#### D. WHERE POSSIBLE, FOSTER VOLUNTARY INFORMATION SHARING COLLABORATIONS

The majority of reported data and public health surveillance systems within this Article were a result of voluntary state and private industry collaborations with the federal government. Currently, many cybersecurity data are voluntarily shared. However, both the scope of, and participation in, these systems is limited. Cybersecurity policy could build out mechanisms illustrated in public health, like the formation of CSTE-like committees of stakeholders, to foster greater community input and collaboration in these systems. Existing information sharing organizations, like ISACs, could serve as foundations for expanding the role and coordinating capacity of stakeholders. Other organizational strategies, such as allowing non-governmental organizations (like those involved in the creation and maintenance of BioSense 2.0) to store and manage data systems may also encourage information sharing. If coupled with the sorts of privacy-sensitive approaches discussed below (particularly law enforcement actions against individuals for non-cybersecurity issues), this can also reduce concerns about the use of shared information for secondary purposes.

E. EMPHASIZE DATA MINIMIZATION, DECENTRALIZATION, AND EARLY DE-IDENTIFICATION

Wherever possible, personally identifiable information should not be collected or shared to support cybersecurity activities. Federal public health reports note that the balance between the need for data sharing and data protection influences how willingly data providers contribute or withhold data.<sup>159</sup> Where personally identifiable data are necessary, they should remain only at the source of collection or intervention. Policies and mechanisms that protect privacy will increase the willingness of entities to share information, and increase the willingness of all stakeholders to consider the potential public cybersecurity benefits of information sharing strategies.

F. PROVIDE ADDITIONAL PRIVACY PROTECTIONS THROUGH NATIONAL INFORMATION SHARING LAWS

The public health system encourages participation by reducing the possibility that information collected for public health purposes will be used to the detriment of individuals. Providing similar protections would build greater acceptance for information sharing. Cybersecurity policies have generally lacked provisions tightly limiting the use of shared information. Their absence has been a major source of objection for civil society stakeholders. Provisions should limit the use of shared information to advancing the public cybersecurity goals of producing better systems and behaviors, manage insecurity, and specifically prohibit the use of information for law enforcement activities that do not directly advance these goals. As in public health, meaningful penalties for violations should accompany these prohibitions and limitations on use. In addition, it would be beneficial to provide protections similar to those afforded by Certificates of Confidentiality, which protect researchers who use public health data from being compelled to release it for legal proceedings. As with public health, sound public cybersecurity policy depends upon ongoing evaluation of the utility of interventions. This research may also involve personally identifiable information, and it too should be protected against disclosure.

---

159. CDC-ATSDR DATA RELEASE GUIDELINES AND PROCEDURES, *supra* note 92, at 5.

G. MAKE AS MUCH CYBERSECURITY DATA AS POSSIBLE OPEN AND ACCESSIBLE FOR PUBLIC USE

As in the field of public health, as much data as possible should be made open and accessible for public use in order to promote public cybersecurity goals. The federal government consolidates and curates public health data and makes it accessible to many stakeholders—from citizen to corporation—through a variety of access and sharing mechanisms.

Data that cannot be made open should be as accessible as possible through limited data access mechanisms and special use agreements. As illustrated in the previous sections, public health takes advantage of several data access modes (public access, access mediated with some restrictions, or no access) to make information accessible for approved purposes. Though data are made as open as possible for public use, great consideration is given to individual privacy and the tradeoffs between accessibility and confidentiality. For public health data offered with some restrictions, the use of RDCs and the ability to run code on sensitive datasets remotely both protect data without inhibiting potential uses.

Data need not be held by the federal government in order to facilitate public access—even for data that requires use and availability restrictions. Data about networked interactions and the state of machines and devices, held and shared only across the private sector, can aid cybersecurity goals. While no single entity has a total view of the data, many have extensive information and insight into the security posture of pieces of the system. It may be far easier, more efficient, and less controversial to bring analysis tools to the data than to bring the data to the government for analysis. While government efforts to advance public cybersecurity goals undoubtedly would benefit from more data, the extent to which the federal government is the appropriate entity to collect it is decidedly unclear. As in public health, multiple models for information collection and access can help balance public cybersecurity goals and other values.

H. CYBERSECURITY SHARING PRACTICES SHOULD EMPHASIZE ETHICAL PUBLIC CYBERSECURITY RESEARCH

Research is essential to many public health goals and is equally important for public cybersecurity. Research evaluates the effectiveness and efficiency of programs, and allows for the formulation of recommendations for improvement. Both preventative and response objectives rely heavily on data and analysis from ongoing research activities. The 1979 Belmont Report guides human subject research activities within the United States, including public health research as it

pertains to interventions or interviews.<sup>160</sup> Built off the canonical Belmont Report, the 2012 Menlo Report<sup>161</sup> establishes an ethical framework for computer and information security research by introducing four core ethical principles, as well as methods to operationalize those principles in the research domain. The core ethical principles include respect for persons, beneficence, justice, and respect for law and the public interest.

Public cybersecurity's primary orientation is focused on society as a whole rather than upon any one individual. But it is vital that data collection and research activities respect individual persons or groups of people who are impacted by data collection, data release, and generalized research findings, or who might ultimately be subjected to containment measures. When promoting public cybersecurity goals, the rights and autonomy of individuals must constantly be factors. Implementing public cybersecurity activities will require tradeoffs between public benefit and individual rights and interests (Table 2). There should also be consideration of how the distribution of the burdens and risks of participation align with the distribution of benefits from public cybersecurity research. For instance, it would not be in the interests of justice to place the administrative burden of information reporting or the loss of privacy disproportionately on one segment of the population unless they were disproportionately to benefit. The selection of subjects within research should be fair, and the burdens should be allocated as equitably as possible so that the risks and benefits are shared among impacted populations. It is imperative that all activities—including information sharing activities—attend to these tensions and tradeoffs, and involve systematic reevaluation of the risks, benefits and burdens as threats evolve.

**Table 2: Tradeoffs Between Data Collection and Surveillance for  
Response Orientation Public Good Activities**

Public Benefit Derived from Data Use	Public Good Activity	Public Interests/Rights
--------------------------------------	----------------------	-------------------------

160. NAT'L COMM'N FOR THE PROTECTION OF HUMAN SUBJECTS OF BIOMEDICAL & BEHAVIORAL RESEARCH, THE BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS RESEARCH (1979), <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>.

161. THE MENLO REPORT: ETHICAL PRINCIPLES GUIDING INFORMATION AND COMMUNICATION TECHNOLOGY RESEARCH (2012), <http://www.dhs.gov/sites/default/files/publications/CSD-MenloPrinciplesCORE-20120803.pdf>.

<ul style="list-style-type: none"> <li>• Manage insecurity from known threats through systematic, organized monitoring</li> <li>• Alerts for unidentified anomalies and new/emerging threats</li> <li>• Immediate contacting of stakeholders affected by detected incident</li> <li>• Ability to trigger other public good activities</li> </ul>	<ul style="list-style-type: none"> <li>• Detection</li> </ul>	<ul style="list-style-type: none"> <li>• Personal autonomy</li> <li>• Individual privacy</li> <li>• Freedom of action</li> <li>• Business interests</li> </ul>
<ul style="list-style-type: none"> <li>• Distinguish between new or recurring threats</li> <li>• Coordinate experts to classify threat or incident</li> <li>• Determine risk and response level</li> <li>• Public announcements about threat/incident</li> </ul>	<ul style="list-style-type: none"> <li>• Identification</li> </ul>	<ul style="list-style-type: none"> <li>• Personal autonomy</li> <li>• Individual privacy</li> <li>• Business interests</li> </ul>
<ul style="list-style-type: none"> <li>• Enable localized and individual action in response to incident</li> <li>• Empower collective action in response to threat</li> <li>• Inform response at many levels to quarantine, patch, or screen for malicious activity</li> <li>• Implement improved preventative techniques to prevent spread to other vulnerable machines and systems</li> </ul>	<ul style="list-style-type: none"> <li>• Containment</li> </ul>	<ul style="list-style-type: none"> <li>• Personal autonomy</li> <li>• Individual privacy</li> <li>• Freedom of action</li> <li>• Business interests</li> <li>• Freedom to innovate</li> </ul>
<ul style="list-style-type: none"> <li>• Appropriately allocate benefits and services to assist recovery</li> <li>• Treat affected and vulnerable populations with patch or design change</li> </ul>	<ul style="list-style-type: none"> <li>• Treatment</li> </ul>	<ul style="list-style-type: none"> <li>• Personal autonomy</li> <li>• Freedom of action</li> <li>• Freedom to innovate</li> </ul>

## VI. CONCLUSION

Information sharing is a means to an end. Its utility must be assessed based on its capacity to support public cybersecurity goals. Orienting cybersecurity policy toward prevention by a reduction in vulnerabilities and

response by managing insecurity, would advance the security of our networks and data. Meeting these objectives will depend on coordinated activities enabled by information. Within public health, information sharing has advanced specific goals and outcomes, in addition to fueling research that has directly and indirectly benefited public health. There are many options for sharing data with different stakeholders and with differing degrees of openness. Laws and institutional policies and practices developed over time in public health provide a rich model that can inform cybersecurity information sharing. This model reflects the need to strike balances between competing public values and the interests of the individual and the collective. The organizational and governance models, policies that address competing values such as privacy, and access mechanisms found in public health provide useful guidance for the development of sound public cybersecurity policy.

## VII. APPENDIX

Public health principles rest heavily on the belief that people are interdependent, which underscores the essence and importance of considering the community. We believe this is also important in the case of cybersecurity, both because networks and systems connect people and data about people, and because there are many communities of practice surrounding cybersecurity.

Table 3: Core Public Health Ethical Principles Applied to Cybersecurity

Principles of the Ethical Practice of Public Health	Application to Practice of Public Cybersecurity
1) Public health should address principally the fundamental causes of disease and requirements for health, aiming to prevent adverse health outcomes.	Public cybersecurity should address systemic design weaknesses and underlying behavioral causes through the preventative orientation to prevent adverse security outcomes.
2) Public health should achieve community health in a way that respects the rights of individuals in the community.	Public cybersecurity should achieve community health in a way that respects the rights of individuals in the community..
3) Public health policies, programs, and priorities should be developed and evaluated through processes that ensure an opportunity for input from community members.	Public cybersecurity policies, programs, and priorities should be developed and evaluated through processes that ensure an opportunity for input from community members.

Principles of the Ethical Practice of Public Health	Application to Practice of Public Cybersecurity
4) Public health should advocate and work for the empowerment of disenfranchised community members, aiming to ensure that the basic resources and conditions necessary for health are accessible to all.	Public cybersecurity should advocate and work for the empowerment of disenfranchised community members (all individual users, all private companies and organizations regardless of size)
5) Public health should seek the information needed to implement effective policies and programs that protect and promote health.	Public cybersecurity should seek the information needed to implement effective policies and programs that protect and promote healthy networks, systems, infrastructure, and use of Internet-based communication.
6) Public health institutions should provide communities with the information they have that is needed for decisions on policies or programs and should obtain the community's consent for their implementation.	Public cybersecurity institutions should provide communities and stakeholders with the information they have that is needed for decisions on policies or programs and should obtain the community and stakeholder's consent for their implementation.
7) Public health institutions should act in a timely manner on the information they have within the resources and the mandate given to them by the public.	Public cybersecurity institutions should act in a timely manner on the information they have within the resources and the mandate given to them by the public..
8) Public health programs and policies should incorporate a variety of approaches that anticipate and respect diverse values, beliefs, and cultures in the community.	Public cybersecurity programs and policies should incorporate a variety of approaches that anticipate and respect diverse values, beliefs, and cultures in the community.
9) Public health programs and policies should be implemented in a manner that most enhances the physical and social environment.	Public cybersecurity programs and policies should be implemented in a manner that most enhances the physical and social environment.
10) Public health institutions should protect the confidentiality of information that can bring harm to an individual or community if made public. Exceptions must be justified on the basis of the high likelihood of significant harm to the individual or others.	Public cybersecurity institutions should protect the confidentiality of information that can bring harm to an individual or community if made public. Exceptions must be justified on the basis of the high likelihood of significant harm to the individual or others.
11) Public health institutions should ensure the professional competence of their employees.	Public cybersecurity institutions should ensure the professional competence of their employees.
12) Public health institutions and their employees should engage in collaborations and affiliations in ways that build the public's trust and the institution's effectiveness.	Public cybersecurity institutions and their employees should engage in collaborations and affiliations in ways that build the public's trust and the institution's effectiveness.

In Table 3, we adapted the key principles within the code of ethics developed by the Public Health Leadership Society to illustrate how they map directly onto the distinctive characteristics found in the doctrine of public cybersecurity.<sup>162</sup> These principles provide guidance during all public good activities and are offered as a way of balancing tensions between collective benefit and individual values, as well as on how to engage various interests of communities and stakeholders. It should be noted that principles 5 through 7 relate specifically to the collection of information, imperative to act upon information, and responsibility to present information to the public. We believe these values support our recommendations on applying public health information sharing mechanisms in the public cybersecurity sphere.

---

162. PUBLIC HEALTH LEADERSHIP SOCIETY, PRINCIPLES OF THE ETHICAL PRACTICE OF PUBLIC HEALTH VERSION 2.2 (2002), <http://phls.org/CMSuploads/Principles-of-the-Ethical-Practice-of-PH-Version-2.2-68496.pdf>.



# CITIZEN SCIENCE: THE LAW AND ETHICS OF PUBLIC ACCESS TO MEDICAL BIG DATA

*Sharona Hoffman*<sup>†</sup>

## ABSTRACT

Patient-related medical information is becoming increasingly available on the Internet, spurred by government open data policies and private sector data sharing initiatives. Websites such as HealthData.gov, GenBank, and PatientsLikeMe allow members of the public to access a wealth of health information. As the medical information terrain quickly changes, the legal system must not lag behind. This Article provides a base on which to build a coherent health data policy. It canvasses emergent data troves and wrestles with their legal and ethical ramifications.

Publicly accessible medical data have the potential to yield numerous benefits, including scientific discoveries, cost savings, new patient support tools, improved healthcare quality, greater government transparency, and public education. At the same time, the availability of electronic personal health information that can be mined by any Internet user raises concerns related to privacy, discrimination, erroneous research findings, and litigation. This Article analyzes the benefits and risks of health data sharing and proposes balanced legislative, regulatory, and policy modifications to guide data disclosure and use.

---

DOI: <http://dx.doi.org/10.15779/Z385Z78>

© 2015 Sharona Hoffman.

<sup>†</sup> Edgar A. Hahn Professor of Law and Professor of Bioethics, Co-Director of Law-Medicine Center, Case Western Reserve University School of Law; B.A., Wellesley College; J.D., Harvard Law School; LL.M. in Health Law, University of Houston. Professor Hoffman was a Distinguished Scholar in Residence at the Centers for Disease Control and Prevention's (CDC) Center for Surveillance, Epidemiology and Laboratory Services during the spring semester of 2014. This Article grew out of the author's work with the CDC, and she wishes to thank the many colleagues who discussed these important issues with her. The author also thanks Jaime Bouvier, Jessie Hill, Tony Moulton, Andy Podgurski, Andrew Pollis, and Timothy Webster for their thoughtful comments on prior drafts. Tracy (Yeheng) Li provided invaluable research assistance throughout this project.

## TABLE OF CONTENTS

I.	INTRODUCTION .....	1744
II.	PUBLICLY AVAILABLE BIG DATA SOURCES .....	1748
A.	FEDERAL AND STATE DATABASES .....	1748
1.	<i>Federal Government Data at HealthData.gov</i> .....	1748
a)	CDC Wonder.....	1749
b)	Chronic Condition Data Warehouse .....	1749
2.	<i>State Government Health Data Websites</i> .....	1750
3.	<i>Healthcare Cost and Utilization Project</i> .....	1750
4.	<i>GenBank</i> .....	1751
5.	<i>All-Payer Claims Databases</i> .....	1752
B.	PRIVATE SECTOR DATABASES .....	1753
1.	<i>Dryad Digital Repository</i> .....	1753
2.	<i>PatientsLikeMe</i> .....	1753
3.	<i>The Personal Genome Project</i> .....	1754
III.	THE BENEFITS OF PUBLIC ACCESS TO HEALTH INFORMATION .....	1755
A.	SCIENTIFIC DISCOVERY .....	1755
B.	RESEARCH COST REDUCTIONS.....	1757
C.	TOOLS TO HELP PATIENTS NAVIGATE THE HEALTHCARE SYSTEM.....	1761
D.	GOVERNMENT TRANSPARENCY AND PUBLIC EDUCATION.....	1761
E.	IMPROVEMENTS IN HEALTHCARE QUALITY AND PUBLIC HEALTH POLICY .....	1762
IV.	RISKS OF PUBLIC ACCESS TO HEALTH DATA .....	1763
A.	PRIVACY THREATS .....	1764
1.	<i>Privacy Law</i> .....	1764
a)	The HIPAA Privacy Rule .....	1765
b)	The Privacy Act.....	1765
c)	State Laws .....	1765
2.	<i>De-identification</i> .....	1766
3.	<i>Does Public-Use Medical Data Pose a Real Privacy Threat?</i> .....	1768
a)	Data Holders Not Covered by the HIPAA Privacy Rule .....	1768
b)	Re-identification of Fully De-identified Health Records.....	1770
c)	The Peculiarities of Genetic Information.....	1771
B.	DISCRIMINATION AND SPECIAL TARGETING.....	1772

1.	<i>Employers</i> .....	1773
a)	Using Identifiable or Re-Identifiable Data .....	1774
b)	De-identified Information as a Basis for Multi-Factor Discrimination and Discrimination by Proxy .....	1776
2.	<i>Financial Institutions and Marketers</i> .....	1778
C.	PROPAGATION OF INCORRECT AND HARMFUL RESEARCH CONCLUSIONS .....	1780
1.	<i>Error Sources</i> .....	1782
2.	<i>Potential Harms</i> .....	1783
D.	LITIGATION.....	1785
1.	<i>Defamation</i> .....	1786
2.	<i>Other Causes of Action</i> .....	1787
3.	<i>Anti-SLAPP Legislation</i> .....	1788
V.	RECOMMENDATIONS .....	1789
A.	PRIVACY AND DATA STEWARDSHIP .....	1790
1.	<i>HIPAA Privacy Rule Modifications</i> .....	1790
a)	Expanding the Definition of “Covered Entity” and Creating National Data Release and De- identification Standards.....	1790
b)	Prohibiting Re-identification .....	1792
2.	<i>Data Release Review Boards</i> .....	1793
3.	<i>Data Use Agreements, Privacy Training, Registries, and Consent Procedures</i> .....	1793
B.	ANTI-DISCRIMINATION PROTECTIONS.....	1796
1.	<i>Detecting, Deterring, and Prosecuting Multi-Factor Discrimination</i> .....	1797
2.	<i>Requiring Disclosure of Data Mining for Disability Proxies and Predictors</i> .....	1798
3.	<i>Addressing Data Mining in the ADA’s Definition of Disability</i> .....	1799
C.	CITIZEN SCIENTIST CHAPERONING .....	1800
D.	TORT CLAIM LITIGATION STRATEGIES.....	1803
VI.	CONCLUSION.....	1804

## I. INTRODUCTION

On May 9, 2013, President Barack Obama issued an executive order entitled “Making Open and Machine Readable the New Default for Government Information.”<sup>1</sup> The Order directed that, to the extent permitted by law, the government must release its data to the public in forms that are easy to find, access, and use.

Health information drawn from patient records is among the most useful but sensitive types of data that are becoming commonly available to the public pursuant to President Obama’s policy and other public and private initiatives that will be discussed in this Article. This is the first article to canvass these emergent data troves and to wrestle with their legal and ethical ramifications. As federal agencies gear up to post increasing amounts of information on the Internet in order to comply with Executive Order 13,642,<sup>2</sup> it is time to carefully consider the benefits and the risks of public access to medical data. The Article also formulates guidelines for data use in order to protect privacy, deter discrimination, and prevent other harms.

Ordinary citizens can now easily find and access patient-related medical data on the Internet.<sup>3</sup> This is the era of “Citizen Science” and “Do-It-Yourself Biology.”<sup>4</sup> Citizen Science is “the practice of public participation and collaboration in scientific research” through data collection, monitoring, and analysis for purposes of scientific discovery,

---

1. Exec. Order No. 13,642, Making Open and Machine Readable the New Default for Government Information, 78 Fed. Reg. 28111 (May 14, 2013), <http://www.gpo.gov/fdsys/pkg/FR-2013-05-14/pdf/2013-11533.pdf>. The Order states, in relevant part:

To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the default state of new and modernized Government information resources shall be open and machine readable. Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable.

2. *Id.*; see also OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB MEMORANDUM M-13-13, OPEN DATA POLICY—MANAGING INFORMATION AS AN ASSET (2013), <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>.

3. See *infra* Part II.

4. Heidi Ledford, *Garage Biotech: Life Hackers*, 467 SCIENCE 650, 650–52 (2010); Amy Dockser Marcus, *Citizen Scientists*, WALL STREET J., Dec. 3, 2011.

usually without compensation.<sup>5</sup> Do-It-Yourself Biology (DIYbio) is an international movement “spreading the use of biotechnology beyond traditional academics and industrial institutions and into the lay public.”<sup>6</sup>

Large data resources are often called “big data,” which is characterized by its sizeable volume, variety, and velocity, that is, the speed with which it is produced.<sup>7</sup> Increasingly, government and private sector sources furnish data collections to the public, and this supply stream will expand considerably in the future.<sup>8</sup> In this Article, publicly available resources will be called “public-use data” or “open data.”

The potential benefits of public access to health information are considerable. At a time of diminishing government funding for research,<sup>9</sup> the widespread availability of high-quality datasets at little to no cost may be very important to continued scientific advancement. Professional researchers as well as talented and dedicated students and amateurs could make important discoveries and answer pressing medical questions,<sup>10</sup> and they could do so without undertaking the expense, time, and work involved in recruiting patients and developing original datasets.<sup>11</sup> Open data has also enabled entrepreneurs to create tools that assist patients in navigating the complexities of the contemporary healthcare system by facilitating searches about symptoms and treatments, listing medical providers by location, and furnishing physician ratings and price information.<sup>12</sup> In addition, federal and state data sharing initiatives promote government transparency and educational initiatives about health and medicine.<sup>13</sup> Finally, data sharing may promote improvements in government-provided services. Easily accessible and navigable public-use

---

5. *Citizen Science*, NAT'L GEOGRAPHIC, <http://education.nationalgeographic.com/education/encyclopedia/citizen-science> (last visited Sept. 17, 2015).

6. DANIEL GRUSHKIN ET AL., SYNTHETIC BIOLOGY PROJECT, SEVEN MYTHS & REALITIES ABOUT DO-IT-YOURSELF BIOLOGY 4 (2013), [http://www.synbioproject.org/process/assets/files/6676/7\\_myths\\_final.pdf](http://www.synbioproject.org/process/assets/files/6676/7_myths_final.pdf).

7. PHILIP RUSSOM, TDWI RESEARCH, BIG DATA ANALYTICS 6 (2011), [https://tdwi.org/research/2011/09/~media/TDWI/TDWI/Research/BPR/2011/TDWI\\_BPRReport\\_Q411\\_Big\\_Data\\_Analytics\\_Web/TDWI\\_BPRReport\\_Q411\\_Big%20Data\\_ExecSummary.pdf](https://tdwi.org/research/2011/09/~media/TDWI/TDWI/Research/BPR/2011/TDWI_BPRReport_Q411_Big_Data_Analytics_Web/TDWI_BPRReport_Q411_Big%20Data_ExecSummary.pdf).

8. *See infra* Part II.

9. Nora Macaluso, *Decade-Long Funding Decline at NIH May Be Poised for Reversal*, *Collins Says*, 13 BLOOMBERG BNA MED. RES. L. & POL'Y REP. 311 (2014) (indicating that “the chances of a project’s getting a grant from NIH have fallen to about 16 percent from 25 percent to 30 percent before 2003”).

10. *See infra* Section III.A.

11. *See infra* note 111 and accompanying text.

12. *See infra* Section III.C.

13. *See infra* Section III.D.

data may help administrators determine how to allocate resources more effectively and engage in quality enhancement activities. Furthermore, media attention focused on healthcare inequities and inefficiencies may catalyze positive policy changes.

At the same time, public access policies are not devoid of risks. First, the possibility of privacy breaches can never be fully eliminated.<sup>14</sup> No matter how carefully data custodians de-identify patient information, at least a small risk of re-identification will always remain. If data holders do not thoroughly anonymize data, the risk of re-identification grows exponentially.<sup>15</sup> Furthermore, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule does not cover most entities that operate public-use databases, and, therefore, those entities are not subject to detailed privacy regulations.<sup>16</sup> Second, open data may enable discrimination by employers, financial institutions, and anyone with a stake in people's health.<sup>17</sup> These entities may attempt to re-identify publicly available health records that belong to applicants or to employees. In the alternative, they may mine medical data to find statistical associations between particular attributes, habits, or behaviors (for example, obesity or smoking) and health risks. Then, based on their findings, entities could formulate discriminatory policies that exclude from employment, financial, or other opportunities individuals they perceive as high-risk.<sup>18</sup> Third, amateurs may reach incorrect conclusions and foster misconceptions among the public about human health or the healthcare industry. Amateurs could disseminate their findings broadly through the Internet without the filter mechanism of having articles reviewed and accepted by peer-reviewed journals.<sup>19</sup> While some errors will be innocent, others might be intentional, with data manipulated to promote personal agendas, such as maligning certain ethnic groups, hurting business competitors, or supporting particular political viewpoints. In turn, parties who believe that they have been hurt by adverse research findings may initiate litigation, asserting claims such as defamation or interference with

---

14. See *infra* Section IV.A.

15. See Sharona Hoffman & Andy Podgurski, *Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research*, 65 SMU L. REV. 85, 105–107 (2012) (discussing re-identification). For further discussion, see *infra* Sections IV.A.2, IV.A.3.b and IV.A.3.c.

16. See *infra* Section IV.A.3.a.

17. See Sharona Hoffman & Andy Podgurski, *In Sickness, Health, and Cyberspace*, 48 B.C. L. REV. 331, 334–35 (2007) (discussing the many parties who might be interested in obtaining medical information about individuals).

18. See *infra* Section IV.B.

19. See *infra* Section IV.C.

economic advantage.<sup>20</sup> In some cases, parties will bring lawsuits merely to intimidate and harass citizen scientists and will needlessly burden the courts.<sup>21</sup>

It is too early to tell whether the benefits of open data will outweigh the risks. However, it is noteworthy that the research projects contemplated in this Article will not be subject to the federal research regulations. The regulations exempt studies based on records or data that are publicly available, and they apply only to studies funded or conducted by federal agencies or submitted to the Food and Drug Administration (FDA) in support of applications for marketing approval.<sup>22</sup> Citizen scientists will therefore operate in a regulatory vacuum with no governing standards or processes for approval and monitoring. This Article argues for the implementation of moderate safeguards and oversight mechanisms that will balance the needs of all stakeholders: patients, researchers, clinicians, industry, federal and state governmental entities, and the public at large.<sup>23</sup>

The Article will proceed as follows. Part II will sample some of the many data collections that various government and private entities have already made publicly available, examining their content and any requirements for data use. Part III will analyze the benefits of public access to medical data, and Part IV will assess its risks. Part V will formulate a detailed proposal for legal and policy interventions designed to promote responsible health data stewardship and to protect those impacted by open data. The first set of recommendations addresses privacy concerns and includes changes to the HIPAA Privacy Rule; establishment of data release review boards; and requirements for data use agreements, privacy training, registries, and consent procedures. Other recommendations in Part V call for clarification and modest expansion of anti-discrimination protections; suggest the development of research guidance, peer review, and publication opportunities for citizen scientists; and address litigation and liability avoidance strategies pertaining to public-use data.

---

20. *See infra* Section IV.D.

21. *See id.*

22. *See* 45 C.F.R. § 46.101(a) (2013) (stating that the regulations apply to “all research involving human subjects conducted, supported or otherwise subject to regulation by any federal department or agency”); 21 C.F.R. § 50.1 (describing the FDA regulations’ scope of coverage); 45 C.F.R. § 46.101(b)(4) (2013) (exempting “[r]esearch involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects”).

23. *See infra* Part V.

## II. PUBLICLY AVAILABLE BIG DATA SOURCES

Many large databases offer public access to patient-related health information. Federal and state governments as well as private sector enterprises have established these databases. No comprehensive catalogue of these sources exists. This Part lists a representative sample of databases that feature public-use medical data.

### A. FEDERAL AND STATE DATABASES

#### 1. *Federal Government Data at HealthData.gov*

HealthData.gov, launched in 2011, is a Department of Health and Human Services website that makes over 1000 data sets available to researchers, entrepreneurs, and the public free of charge.<sup>24</sup> It predates Executive Order 13,642 by two years and establishes a home for the federal government's open data. Several states and federal government agencies such as the Centers for Disease Control and Prevention (CDC), the Centers for Medicare & Medicaid Services (CMS), the National Institutes of Health (NIH), and the Administration for Children and Families provide the data sets.<sup>25</sup>

All users can search for information by key words, agency type, and subject area.<sup>26</sup> As just one example, users can access a table entitled "Vaccination coverage among children 19–35 months of age for selected diseases, by race, Hispanic origin, poverty level, and location of residence in metropolitan statistical area."<sup>27</sup> HealthData.gov offers many interactive analysis tools and will continue to grow and be refined over the coming years.<sup>28</sup> Users can access a number of separate federal agency databases

---

24. *About*, HEALTHDATA.GOV, [www.healthdata.gov/content/about](http://www.healthdata.gov/content/about) (last visited Nov. 23, 2015).

25. *Id.*

26. HEALTHDATA.GOV, <http://healthdata.gov> (last visited Nov. 23, 2015). The subject areas listed are administrative, biomedical research, children's health, epidemiology, healthcare cost, healthcare providers, Medicaid, Medicare, population statistics, quality measurement, safety, treatments, and other.

27. Ctrs. for Disease Control & Prevention, *Vaccination Coverage Among Children 19–35 Months of Age for Selected Diseases*, HEALTHDATA.GOV (Oct. 14, 2015), <http://www.healthdata.gov/dataset/selected-trend-table-health-united-states-2011-vaccination-coverage-among-children-19-35>.

28. *See, e.g.*, Harnam Singh, *The National Practitioner Data Bank (NPDB) Introduces Interactive Data Analysis Applications*, HEALTHDATA.GOV (May 29, 2014), <http://healthdata.gov/blog/national-practitioner-data-bank-npdb-introduces-interactive-data-analysis-applications>; Damon Davis et al., *Health Data Initiative Strategy & Execution Plan Released and Ready for Feedback*, HEALTHDATA.GOV (Oct. 23, 2013), <http://www.healthdata.gov/blog/health-data-initiative-strategy-execution-plan-released-and-ready-feedback>.

through Healthdata.gov. The CDC database, CDC Wonder,<sup>29</sup> and the CMS database, Chronic Condition Data Warehouse,<sup>30</sup> are discussed below.

a) CDC Wonder

CDC Wonder enables researchers and the public at large to access a wide variety of public health information.<sup>31</sup> This includes data sets about deaths, births, cancer, HIV and AIDS, tuberculosis, vaccinations, census data, and more.<sup>32</sup> The website features statistical research data, reference material, reports, and guidelines related to public health.<sup>33</sup> Users conduct queries by selecting items from drop-down menus and completing fill-in-the-blank forms.<sup>34</sup> Prior to receiving data, users must read a short “data use restrictions” screen and click “I agree,” thereby promising to comply with instructions concerning data use and disclosure that are designed to protect the privacy of data subjects.<sup>35</sup>

b) Chronic Condition Data Warehouse

The CMS established the Chronic Condition Data Warehouse (CCW) to allow users to purchase data about Medicare and Medicaid beneficiaries and claims.<sup>36</sup> Researchers can apply for access to identifiable or partially identifiable data, and CCW administrators scrutinize all requests.<sup>37</sup> CCW also offers public-use files that contain aggregated summary level health information for which no data use agreement or

---

29. See Ctrs. for Disease Control & Prevention, *CDC Wonder: Births*, HEALTHDATA.GOV (Oct. 30, 2015), <http://healthdata.gov/dataset/cdc-wonder-births-0>.

30. Dep't of Health & Human Servs., *Chronic Condition Data Warehouse*, HEALTHDATA.GOV (Oct. 30, 2015), <http://www.healthdata.gov/dataset/chronic-condition-data-warehouse>.

31. *What Is CDC Wonder?*, CDC WONDER, [http://wonder.cdc.gov/wonder/help/main.html#What is WONDER](http://wonder.cdc.gov/wonder/help/main.html#What%20is%20WONDER) (last updated Jan. 25, 2016).

32. *Id.*

33. *Id.*

34. *Id.*

35. See, e.g., *About Natality, 2007–2013*, CDC WONDER, <http://wonder.cdc.gov/natality-current.html> (last visited Nov. 23, 2015). See *infra* note 309 and accompanying text for further discussion of data use agreements.

36. CENTERS FOR MEDICARE & MEDICAID SERVICES, CHRONIC CONDITIONS DATA WAREHOUSE, <https://www.ccwdata.org/web/guest/home> (last visited Nov. 23, 2015).

37. *CMS Data Request Center*, RESEARCH DATA ASSISTANCE CENTER, <http://www.resdac.org/cms-data/request/cms-data-request-center> (last visited Nov. 23, 2015).

privacy board review is required.<sup>38</sup> For example, the Medicaid State Drug Utilization File contains information about outpatient drugs for which state Medicaid agencies have paid.<sup>39</sup>

### 2. *State Government Health Data Websites*

Like the federal government, many states offer publicly available health data on government websites. Examples are Health Data New York,<sup>40</sup> New Jersey State Health Assessment Data,<sup>41</sup> North Carolina State Center for Health Statistics,<sup>42</sup> FloridaHealthFinder.gov,<sup>43</sup> and Minnesota Center for Health Statistics.<sup>44</sup> All these websites provide a wealth of information free of charge to the public and offer a variety of interactive tools and query options.

### 3. *Healthcare Cost and Utilization Project*

The Healthcare Cost and Utilization Project (HCUP) is sponsored by the Agency for Healthcare Research and Quality<sup>45</sup> and offers a variety of databases for purchase. These include the following:

- Nationwide Inpatient Sample
- Kids' Inpatient Database
- Nationwide Emergency Department Sample
- State Inpatient Databases
- State Ambulatory Surgery Databases
- State Emergency Department Databases<sup>46</sup>

38. *Public Use Files (PUF)/Non-Identifiable Data Requests*, RESEARCH DATA ASSISTANCE CENTER, <http://www.resdac.org/cms-data/request/public-use-files> (last visited Nov. 23, 2015).

39. *Medicaid State Drug Utilization File*, RESEARCH DATA ASSISTANCE CENTER, <http://resdac.advantagelabs.com/cms-data/files/medicaid-state-drug-utilization> (last visited Nov. 23, 2015).

40. *Health Data NY*, N.Y. DEP'T OF HEALTH, <https://health.data.ny.gov> (last visited Nov. 23, 2015).

41. *NJSHAD: New Jersey's Public Health Data Resource*, N.J. DEP'T OF HEALTH, <https://www26.state.nj.us/doh-shad/home/Welcome.html> (last updated Jan. 5, 2016).

42. *Statistics and Reports*, N.C. STATE CEN. FOR HEALTH STATISTICS, <http://www.schs.state.nc.us/data/minority.cfm> (last updated Jan. 5, 2016).

43. *State Health Data Directory*, FLA. AGENCY FOR HEALTH CARE ADMIN., <http://www.floridahealthfinder.gov/StateHealthDataDirectory> (last visited Nov. 23, 2015).

44. *Selected Public Health Data Websites*, MINN. CENTER FOR HEALTH STAT., <http://www.health.state.mn.us/divs/chs/countytables/resources.htm> (last updated Jan. 21, 2016).

45. *Overview of HCUP*, HEALTHCARE COST & UTILIZATION PROJECT, <http://www.hcup-us.ahrq.gov/overview.jsp> (last updated Jan. 28, 2016).

46. *Id.*

HCUP databases offer “a core set of clinical and nonclinical information found in a typical [hospital] discharge abstract including all-listed diagnoses and procedures, discharge status, patient demographics, and charges for all patients, regardless of payer (e.g., Medicare, Medicaid, private insurance, uninsured).”<sup>47</sup> Patient demographics may include sex, age, and—for some states—race, but no other attributes that more directly identify patients.<sup>48</sup> The databases are available for purchase, and purchasers must complete a training course and sign a data use agreement prior to receiving data.<sup>49</sup> Users must agree to use the data solely for research and statistical purposes and not to attempt to identify any individual.<sup>50</sup> Those wishing to purchase information from state databases must also explain how they intend to use the data.<sup>51</sup> Prices may vary significantly, depending on the type of data sought and the type of entity with which the applicant is affiliated (for example, for-profit or non-profit organization), with significant discounts available to students.<sup>52</sup>

#### 4. *GenBank*

GenBank is the National Institutes of Health’s genetic sequence database, which includes all DNA sequences that are publicly available.<sup>53</sup> The data are free, and GenBank places no restriction on their use.<sup>54</sup> According to scientists at the National Center for Biotechnology Information, GenBank contains “over 900 complete genomes, including the draft human genome, and some 95,000 species.”<sup>55</sup> Leading journals

---

47. *Databases and Related Tools from HCUP: Fact Sheet*, AGENCY FOR HEALTHCARE RESEARCH & QUALITY, <http://archive.ahrq.gov/research/findings/factsheets/tools/hcupdata/datahcup.html> (last updated Mar. 2011).

48. *Overview of the State Inpatient Databases*, HEALTHCARE COST & UTILIZATION PROJECT, <http://www.hcup-us.ahrq.gov/sidoverview.jsp> (last updated Jan. 20, 2016).

49. *Purchase HCUP Data*, HEALTHCARE COST & UTILIZATION PROJECT, [http://www.hcup-us.ahrq.gov/tech\\_assist/centdist.jsp](http://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp) (last updated Nov. 18, 2015).

50. HEALTHCARE COST & UTILIZATION PROJECT, HCUP NATIONWIDE INPATIENT SAMPLE APPLICATION (2015), [http://www.hcup-us.ahrq.gov/db/nation/nis/NISApp\\_Final.pdf](http://www.hcup-us.ahrq.gov/db/nation/nis/NISApp_Final.pdf).

51. *Purchase HCUP Data*, *supra* note 49.

52. HEALTHCARE COST & UTILIZATION PROJECT, SID/SASD/SEDD APPLICATION KIT (2015), [http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD\\_Final.pdf](http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf) (listing prices that range from \$35 to over \$1600).

53. *GenBank Overview*, NAT’L CEN. FOR BIOTECH. INFO. (NCBI), <http://www.ncbi.nlm.nih.gov/genbank> (last visited Nov. 23, 2015).

54. *Id.*

55. Jo McEntyre & David J. Lipman, *GenBank—A Model Community Resource?*, NATURE (Apr. 5, 2001), <http://www.nature.com/nature/debates/e-access/Articles/lipman.html>.

now require authors to deposit their sequences in GenBank, and all publicly funded laboratories also do so as a matter of policy.<sup>56</sup>

GenBank provides a variety of data search and retrieval tools, such as the Basic Local Alignment Search Tool (BLAST), which finds similarities between sequences.<sup>57</sup> Public-use data available on GenBank have enabled scientists and commercial enterprises to conduct research and generate new products, including assemblies of the human genome produced by Celera Genomics and the University of California at Santa Cruz.<sup>58</sup>

### 5. *All-Payer Claims Databases*

A large number of states have launched all-payer claims databases that collect information about private and public insurance related to medical, dental, and pharmacy services.<sup>59</sup> Typically, the collected data include information regarding patient demographics; insurance contracts; healthcare providers; payments made by insurers and patients; dates on which medical services were received; and codes for diagnoses, procedures, and drugs.<sup>60</sup> Consumers, employers, and other stakeholders can access data to learn about healthcare costs, compare prices, and make more informed decisions about insurance plans and healthcare providers.<sup>61</sup>

Similarly, CMS has released Medicare provider utilization and payment data that is available free of charge.<sup>62</sup> The website offers information pertaining to the 100 most commonly performed inpatient services, thirty frequently provided outpatient services, and more.<sup>63</sup> Thus, for instance, users may obtain hospital-specific charges for particular services and compare prices.<sup>64</sup>

---

56. *Id.*

57. *Id.*; *Genbank Overview*, *supra* note 53.

58. McEntyre & Lipman, *supra* note 55.

59. JO PORTER ET AL., THE BASICS OF ALL-PAYER CLAIMS DATABASES: A PRIMER FOR STATES 1 (2014), <http://www.apcdouncil.org/sites/apcdouncil.org/files/The%20Basics%20of%20All-Payer%20Claims%20Databases.pdf>.

60. *Id.* at 2.

61. *Id.* at 3; *Colorado Medical Price Compare*, CTR. FOR IMPROVING VALUE IN HEALTH CARE, <https://www.cohealthdata.org> (last visited Nov. 23, 2015); *CHIA Data*, CTR. FOR HEALTH INFO. & ANALYSIS, <http://www.chiamass.gov/chia-data> (last visited Nov. 23, 2015) (requiring applications for Massachusetts data).

62. *Medicare Provider Utilization and Payment Data*, CTRS. FOR MEDICARE & MEDICAID SERVS., <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data> (last updated Apr. 30, 2015).

63. *Id.*

64. *Medicare Provider Utilization and Payment Data: Inpatient*, CTRS. FOR MEDICARE & MEDICAID SERVS., <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient.html>

## B. PRIVATE SECTOR DATABASES

### 1. *Dryad Digital Repository*

Dryad is an international repository containing data files associated with peer-reviewed scientific articles and other “reputable sources (such as dissertations).”<sup>65</sup> It is a nonprofit organization supported by fees from its members and data submitters.<sup>66</sup> Researchers submit data underlying their publications directly to Dryad, and any member of the public can access the collection at no cost.<sup>67</sup> The website provides a search tool that allows users to enter key words or other search criteria and takes them to data associated with particular publications.<sup>68</sup>

### 2. *PatientsLikeMe*

PatientsLikeMe is a for-profit website that enables patients who sign up for membership to share their health data and disease experiences.<sup>69</sup> Users can report and obtain information about treatments and connect with others who have the same condition.<sup>70</sup> PatientsLikeMe acknowledges that it sells de-identified information submitted by users to its “partners,” which it describes as “companies that can use that data to improve or understand products or the disease market.”<sup>71</sup> Members may choose different privacy settings and may determine whether non-members will

---

(last updated June 1, 2015). *But see* Patrick T. O’Gara, *Caution Advised: Medicare’s Physician-Payment Data Release*, 371 NEW ENG. J. MED. 101 (2014) (discussing the limitations of payment data released by CMS); Dawn Fallik, *For Big Data, Big Questions Remain*, 33 HEALTH AFF. 1111, 1111 (2014) (stating that “Medicare’s release of practitioner payments highlights the strengths and weaknesses of digging into big data”).

65. *The Organization: Overview*, DRYAD, <http://datadryad.org/pages/organization> (last updated Oct. 22, 2015); *Frequently Asked Questions*, DRYAD, <http://datadryad.org/pages/faq#depositing> (last updated Jan. 5, 2016).

66. *Pricing Plans and Data Publishing Prices*, DRYAD, <http://datadryad.org/pages/pricing> (last updated Jan. 5, 2016).

67. *Frequently Asked Questions*, DRYAD, <http://datadryad.org/pages/faq#using> (last updated Jan. 5, 2016).

68. *The Repository: Key Features*, DRYAD, <http://datadryad.org/pages/repository> (last updated Feb. 15, 2015).

69. PATIENTSLIKEME, <http://www.patientslikeme.com> (last visited Nov. 23, 2015).

70. *What Is PatientsLikeMe?*, PATIENTSLIKEME, <https://support.patientslikeme.com/hc/en-us/articles/201186434-What-is-PatientsLikeMe-> (last visited Nov. 23, 2015).

71. *Does PatientsLikeMe Sell My Data?*, PATIENTSLIKEME, <https://support.patientslikeme.com/hc/en-us/articles/201245770-Does-PatientsLikeMe-sell-my-information-> (last visited Nov. 23, 2015).

be able to view any of their data.<sup>72</sup> PatientsLikeMe releases reports of aggregated data concerning symptoms and treatments to the public.<sup>73</sup> In addition, members may opt into a public registry that will make their profiles and shared data available to anyone with access to the Internet.<sup>74</sup> PatientsLikeMe makes public-use information available on its website at no cost and does not require applications or data use agreements.<sup>75</sup>

### 3. *The Personal Genome Project*

George Church launched the Personal Genome Project in 2005 at Harvard University, and it is now an international enterprise involving thousands of patients.<sup>76</sup> It aims to promote research and offers genomic, environmental, and human trait information from volunteer participants to any interested party.<sup>77</sup> Users can easily access a wealth of information directly from the website, including genome data, genome reports, trait and survey data, participant profiles, and microbiome data.<sup>78</sup> Data files list the date of birth, gender, zip code, height, weight, and race of individual participants, though names are not displayed.<sup>79</sup> The Personal Genome Project states explicitly that its participants must be “willing to waive expectations of privacy” in order to make “a valuable and lasting contribution to science.”<sup>80</sup>

---

72. *Privacy Policy*, PATIENTSLIKEME, <http://www.patientslikeme.com/about/privacy> (last updated Mar. 5, 2012).

73. *See, e.g., Treatments*, PATIENTSLIKEME, <http://www.patientslikeme.com/treatments> (last updated Feb. 2, 2016).

74. *See, e.g., Welcome to the PatientsLikeMe Public ALS Registry*, PATIENTSLIKEME, <http://www.patientslikeme.com/registry> (last visited Nov. 23, 2015); *What Information is Visible on Public Profiles?*, PATIENTSLIKEME, <https://support.patientslikeme.com/hc/en-us/articles/201245830-What-information-is-visible-on-public-profiles-> (last visited Nov. 23, 2015).

75. *Conditions at PatientsLikeMe*, PATIENTSLIKEME, <http://www.patientslikeme.com/conditions> (last updated Feb. 5, 2016).

76. *About PGP Harvard*, PERSONAL GENOME PROJECT, HARV. MED. SCH., <http://www.personalgenomes.org/harvard/about-pgp> (last visited Nov. 23, 2015).

77. *Id.*

78. *Id.; Data & Samples*, PERSONAL GENOME PROJECT, HARV. MED. SCH., <http://www.personalgenomes.org/harvard/data> (last visited Nov. 23, 2015) (Microbiome data focuses on “the types of bacteria in and on a participant’s body.”).

79. *See, e.g., Public Genetic Data*, PERSONAL GENOME PROJECT, HARV. MED. SCH., [https://my.pgp-hms.org/public\\_genetic\\_data](https://my.pgp-hms.org/public_genetic_data) (last visited Nov. 23, 2015).

80. *About PGP Harvard*, *supra* note 76.

### III. THE BENEFITS OF PUBLIC ACCESS TO HEALTH INFORMATION

Public-use data potentially offer many valuable benefits. These include new scientific discoveries, research cost savings, new tools to help patients navigate the healthcare system, greater government transparency, public education about science and medicine, improved healthcare quality, and positive healthcare policy changes.

#### A. SCIENTIFIC DISCOVERY

One of the great hopes of health data sharing is that it will promote scientific discovery and medical advances. Citizen scientists may be extremely motivated and dedicated researchers, perhaps especially if they are focusing on diseases that afflict them or their loved ones. Citizen scientists who would not otherwise have access to health data and lack the means to collect original data for studies may nevertheless have the skills, talent, and creativity to make significant contributions given the appropriate data tools.<sup>81</sup>

In his May 2013 executive order, President Obama stated that public information resources have enabled entrepreneurs and innovators “to develop a vast range of useful new products and businesses.”<sup>82</sup> Similarly, proponents of DIYbio enthuse that it “can inspire a generation of bioengineers to discover new medicines, customize crops to feed the world’s exploding population, harness microbes to sequester carbon, solve the energy crisis, or even grow our next building materials.”<sup>83</sup>

Citizen scientists have proven themselves to be capable inventors whose contributions aid many people. For example, three Dutch DIY biologists created Amplino, an inexpensive diagnostic system that can be used in developing countries to detect malaria with a single drop of blood in less than forty minutes.<sup>84</sup> Likewise, Katherine Aull, a graduate of the Massachusetts Institute of Technology whose father suffered from

---

81. Huseyin Naci & John P. A. Ioannidis, *Evaluation of Wellness Determinants and Interventions by Citizen Scientists*, 314 JAMA 121, 122 (2015), <http://jama.jamanetwork.com/article.aspx?articleid=2330497>.

82. Exec. Order No. 13,642, *supra* note 1.

83. Grushkin et al., *supra* note 6, at 4.

84. Thomas Landrain et al., *Do-It-Yourself Biology: Challenges and Promises for an Open Science and Technology Movement*, 7 SYST. SYNTHETIC BIOLOGY 115, 121 (2013); Linda Nordling, *DIY Biotech: How to Build Yourself a Low-Cost Malaria Detector*, GUARDIAN (Apr. 25, 2014), <http://www.theguardian.com/global-development-professionals-network/2014/apr/25/diy-detector-malaria-eradication-amplino> (reporting that Amplino “is almost ready for field-testing in rural Zambia”).

hemochromatosis, a condition that causes the body to absorb excessive amounts of iron and can permanently damage vital organs, developed a homemade genetic test to determine whether she was vulnerable to this inherited disease.<sup>85</sup> She built a lab in her closet and used equipment purchased from eBay or found in her kitchen.<sup>86</sup>

New troves of publicly available data promise to facilitate and accelerate the work of professional researchers and citizen scientists. Public data sources have already led to important discoveries. For example, Project Tycho is a University of Pittsburgh initiative designed to promote the availability and use of public health data by facilitating its analysis and redistribution.<sup>87</sup> Tycho researchers have digitized disease surveillance data from the years 1888 to 2011 published in the CDC's *Morbidity and Mortality Weekly Report* and estimate that since 1924, 103 million incidents of childhood diseases were prevented because of immunizations.<sup>88</sup> This finding will be useful for public health authorities, who at times meet resistance to vaccination efforts.

Among the more creative initiatives was a crowdsourcing contest called the Dialogue on Reverse Engineering Assessment and Methods (DREAM7) focused on breast cancer prognosis.<sup>89</sup> Crowdsourcing can be defined as "a participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals . . . via a flexible open call, the voluntary undertaking of a task."<sup>90</sup> DREAM7 provided participants with access to genetic and clinical data from Sage's Synapse, an informatics platform that allows users to

---

85. Ana Delgado, *DIYbio: Making Things and Making Futures*, 48 *FUTURES* 65, 70 (2013); *Biopunks Tinker with the Building Blocks of Life*, NPR (May 19, 2011), <http://www.npr.org/2011/05/22/136464041/biopunks-tinker-with-the-building-blocks-of-life>.

86. Delgado, *supra* note 85, at 70.

87. *About Project Tycho Data*, U. OF PITTSBURGH, <https://www.tycho.pitt.edu/about.php> (last visited Nov. 23, 2015).

88. Willem G. van Panhuis et al., *Contagious Diseases in the United States from 1888 to the Present*, 369 *NEW ENG. J. MED.* 2152, 2156 (2013).

89. Michael Eisenstein, *Crowdsourced Contest Identifies Best-In-Class Breast Cancer Prognostic*, 7 *NATURE BIOTECH.* 578, 578 (2013).

90. Enrique Estellés-Arolas & Fernando González-Ladrón-de-Guevara, *Towards an Integrated Crowdsourcing Definition*, 38 *J. INFO. SCI.* 189, 197 (2012); *see also* Thea C Norman et al., *Leveraging Crowdsourcing to Facilitate the Discovery of New Medicines*, 3 *SCI. TRANSLATIONAL MED.* mr1, 2 (2011) (defining crowdsourcing as "the act of outsourcing tasks traditionally performed by an employee to an undefined, large group of people or community (a 'crowd')").

share data and access programming codes and analytical tools.<sup>91</sup> The contest challenged the crowd to “provide an unbiased assessment of models and methodologies for the prediction of breast cancer survival.”<sup>92</sup> A winner was selected from among 1400 entries, and results were published in a scientific journal.<sup>93</sup>

Crowdsourcing is an increasingly popular phenomenon.<sup>94</sup> It has been used for projects ranging from locating over 1400 automated external defibrillators in public places in Philadelphia to developing a predictive algorithm for regions of local similarity between genetic sequences that is superior to the NIH’s standard algorithm, BLAST.<sup>95</sup> The availability of vast amounts of publicly accessible data may make crowdsourcing all the more prevalent. Researchers will likely continue to harness the talents and expertise of citizen scientists to make important contributions to medical science.<sup>96</sup>

## B. RESEARCH COST REDUCTIONS

Open data resources will be of particular value in an era of diminished research funding. NIH appropriations peaked at \$36.4 billion in fiscal year 2010 thanks to funding from the American Recovery and Reinvestment Act, but they declined to \$29.9 billion by fiscal year 2014. In 2014, the NIH funded 18.1% of grant proposals compared to 31.5% in 2000.<sup>97</sup>

At the same time, despite the abundance of information and medical technology available in the twenty-first century, “more than half of medical treatments are used without sufficient proof of their

---

91. SYNAPSE, ABOUT SYNAPSE (2013), [https://s3.amazonaws.com/static.synapse.org/About\\_Synapse.pdf](https://s3.amazonaws.com/static.synapse.org/About_Synapse.pdf).

92. GUSTAVO STOLOVITZSKY & ANDREA CALIFANO, *DREAM CHALLENGE* (2013), <http://www.slideshare.net/tulipnandu/dream-challenge>.

93. Eisenstein, *supra* note 89, at 578.

94. Benjamin M. Good & Andrew I. Su, *Crowdsourcing for Bioinformatics*, 29 *BIOINFORMATICS* 1925, 1925 (2013).

95. *The Accelerating World of Drug Discovery and Commercialization*, *TRENDS MAG.*, Oct. 2013, at 30 (2013); *Basic Local Assignment Search Tool (BLAST)*, NAT’L CENTER FOR BIOTECHNOLOGY INFORMATION, <http://blast.ncbi.nlm.nih.gov/Blast.cgi> (last visited Nov. 23, 2015).

96. Benjamin L. Raynard et al., *Crowdsourcing—Harnessing the Masses to Advance Health and Medicine, A Systematic Review*, 29 *J. GEN. INTERNAL MED.* 187, 187 (2014) (concluding that “[u]tilizing crowdsourcing can improve the quality, cost, and speed of a research project while engaging large segments of the public and creating novel science”).

97. *Research Project Success Rates by NIH Institute for 2014*, U.S. DEP’T OF HEALTH & HUMAN SERVICES, [http://www.report.nih.gov/success\\_rates/Success\\_ByIC.cfm](http://www.report.nih.gov/success_rates/Success_ByIC.cfm) (last updated Mar. 22, 2012).

effectiveness.”<sup>98</sup> For example, experts have recently raised new questions about the efficacy of mammography, a well-established practice that was long considered life-saving and a key element of preventive medicine.<sup>99</sup> Likewise, although physicians have prescribed and studied hormone replacement therapy for post-menopausal women for decades, experts are still unsure as to whether it is advisable or whether its risks outweigh its benefits, at least for some subgroups of patients.<sup>100</sup> A third illustration is a debate over the risks of a particular class of antidepressants called selective serotonin reuptake inhibitors (SSRIs) in light of evidence that they may induce suicidal thoughts and behavior in adolescent patients.<sup>101</sup> No consensus has formed regarding this side effect, and further study is necessary.<sup>102</sup>

Professional researchers and citizen scientists will be able to use open data to reduce the expense of clinical trials and to conduct low-cost records-based research. While many will focus on well-known and widespread health problems, open data may also stimulate the study of subjects for which little to no public funding is available. For example, because of vigorous lobbying by the National Rifle Association, the CDC was prohibited for many years from analyzing the impact of firearms on public health.<sup>103</sup> Similarly, there is often limited interest in or funding for research relating to rare diseases.<sup>104</sup> Citizen scientists, however, may be highly motivated, for personal rather than profit-seeking reasons, to research those diseases.

---

98. Eric B. Larson, *Building Trust in the Power of “Big Data” Research to Serve the Public Good*, 309 JAMA 2443, 2444 (2013).

99. Nikola Biller-Andorno & Peter Jüni, *Abolishing Mammography Screening Programs? A View from the Swiss Medical Board*, 370 NEW ENG. J. MED. 1965, 1965–67 (2014).

100. HERBERT I. WEISBERG, *BIAS AND CAUSATION: MODELS AND JUDGMENT FOR VALID COMPARISONS* 18–21 (2010) (noting that the risks may include slight elevations in the incidence of coronary heart disease and breast cancer).

101. *Id.* at 21–23.

102. *Id.*

103. Michael Luo, *Sway of N.R.A. Blocks Studies, Scientists Say*, N.Y. TIMES (Jan. 26, 2011), <http://www.nytimes.com/2011/01/26/us/26guns.html>. The moratorium was lifted by a memorandum issued by President Obama in January of 2013. Memorandum, *Engaging in Public Health Research on the Causes and Prevention of Gun Violence*, 78 Fed. Reg. 4295 (Jan. 16, 2013).

104. NAT’L ORG. FOR RARE DISORDERS, *RESEARCH GRANT POLICY* (2015), <https://rarediseases.org/wp-content/uploads/2015/05/NORD-Research-Grant-Policy.pdf>.

The gold standard of medical research has traditionally been randomized, controlled clinical trials.<sup>105</sup> Phase 3 clinical trials, conducted as the final step before approval of a drug, cost an average of \$20 million, involve 300 to 3000 people, and last one to four years.<sup>106</sup> These experimental studies are conducted through “the collection of data on a process when there is some manipulation of variables that are assumed to affect the outcome of a process, keeping other variables constant as far as possible.”<sup>107</sup> Thus, investigators might design a clinical trial to compare two drugs for a particular ailment or to compare a drug to a placebo. If researchers share data from prior clinical trials, they may be able to improve study quality and efficiency by honing in on patient sub-groups that are most likely to be responsive to the drug in question.<sup>108</sup> For example, a bladder cancer study determined that one participant who responded unusually well to the drug everolimus had a particular genetic mutation, and thus future testing of the drug could focus on subjects with that mutation to determine whether it enhances responsiveness to the drug.<sup>109</sup>

In the alternative, researchers can undertake observational studies by reviewing existing records and data sets rather than conducting experiments.<sup>110</sup> Professional researchers and citizen scientists will be able to use the large quantities of open data that are now becoming available to

---

105. Friedrich K. Port, *Role of Observational Studies Versus Clinical Trials in ESRD Research*, 57 KIDNEY INT'L (SUPPLEMENT 74) S3, S3 (2000), [http://www.kidney-international.org/article/S0085-2538\(15\)47033-4/pdf](http://www.kidney-international.org/article/S0085-2538(15)47033-4/pdf) (stating that “[r]andomized controlled clinical trials have been considered by many to be the only reliable source for information in health services research”); see also Sharon Hoffman, *The Use of Placebos in Clinical Trials: Responsible Research or Unethical Practice?*, 33 CONN. L. REV. 449, 452–54 (2001) (describing different clinical trial designs).

106. U.S. DEP'T OF HEALTH & HUMAN SERVICES, OFFICE OF THE ASSISTANT SECRETARY FOR PLANNING AND EVALUATION, EXAMINATION OF CLINICAL TRIAL COSTS AND BARRIERS FOR DRUG DEVELOPMENT (2014), <http://aspe.hhs.gov/report/examination-clinical-trial-costs-and-barriers-drug-development>; *Step 3: Clinical Research*, U.S. FDA, (2015), <http://www.fda.gov/ForPatients/Approvals/Drugs/ucm405622.htm>. These sources also discuss the earlier stages of clinical trials, Phase 1 and Phase 2.

107. BRYAN F.J. MANLY, *THE DESIGN AND ANALYSIS OF RESEARCH STUDIES* 1 (1992).

108. Eisenstein, *supra* note 89, at 580.

109. *Id.*; Gopa Iyer et al., *Genome Sequencing Identifies a Basis for Everolimus Sensitivity*, 338 SCIENCE 221, 221 (2012).

110. Observational studies involve the review of existing records or data. See CHARLES P. FRIEDMAN & JEREMY C. WYATT, *EVALUATION METHODS IN BIOMEDICAL INFORMATICS* 369 (Kathryn J. Hannah & Marion J. Ball eds., 2nd ed. 2006) (defining observational studies as involving an “[a]pproach to study design that entails no experimental manipulation”).

minimize research expenses. Researchers may find that existing data collections contain all the raw data that they need and be spared the work and cost of recruiting human subjects for original data. Public-use data can thus prevent costly duplication of effort.<sup>111</sup>

Furthermore, an emerging trend called crowdfunding can fund relatively inexpensive big data projects.<sup>112</sup> Crowdfunding is an Internet-based method of fundraising by which one can solicit money from numerous donors, who usually contribute small amounts.<sup>113</sup> Typically, crowdfunding for scientific projects raises less than \$10,000,<sup>114</sup> but enterprising fund-raisers have frequently surpassed that sum.<sup>115</sup> Public-use data may enable a growing number of projects to have very limited costs that researchers can cover creatively rather than through the traditional channels of government-allocated grant awards.

---

111. CDC, CDC-GA-2005-14, CDC/ATSDR POLICY ON RELEASING AND SHARING DATA 5–6 (2005), <http://www.cdc.gov/maso/Policy/ReleasingData.pdf>.

112. Vural Özdemir et al., *Crowd-Funded Micro-Grants for Genomics and “Big Data”*: An Actionable Idea Connecting Small (Artisan) Science, Infrastructure Science, and Citizen Philanthropy, 17 OMICS 161, 162 (2013).

113. Stuart R. Cohn, *New Crowdfunding Registration Exemption: Good Idea, Bad Execution*, 64 FLA. L. REV. 1433, 1434 (2012).

114. Rachel E. Wheat et al., *Raising Money for Scientific Research Through Crowdfunding*, 28 TRENDS ECOLOGY & EVOLUTION 71, 72 (2013), [http://jarrettbyrnes.info/pdfs/Wheat\\_et\\_al\\_2012.pdf](http://jarrettbyrnes.info/pdfs/Wheat_et_al_2012.pdf).

115. Ethan O. Perlstein, *Anatomy of the Crowd4Discovery Crowdfunding Campaign*, 2 SPRINGERPLUS 560, 561 (2013), <http://www.springerplus.com/content/pdf/2193-1801-2-560.pdf> (reporting that the authors raised \$25,460 from 390 donors in 15 countries for a pharmacological research project); Joe Palca, *Scientists Get Research Donations from Crowd Funding*, NPR (Mar. 15, 2013), <http://www.npr.org/2013/02/14/171975368/scientist-gets-research-donations-from-crowdfunding> (reporting that UBiome and American Gut together raised over \$600,000 for projects designed to discover how microbiomes (tiny organisms that reside in the human body) influence health when donors were promised an analysis of the bacteria in their own digestive tracts). The Internet offers a large number of platforms for crowdfunding, including the aptly named Kickstarter, Experiment, and Indiegogo, among others. See KICKSTARTER, <https://www.kickstarter.com>; EXPERIMENT, <https://experiment.com>; INDIEGOGO, <https://www.indiegogo.com>. Crowdfunding has become so popular that it is being used not only by enterprising individuals and companies but also by several universities, such as the University of Virginia and Tulane, that are seeking to compensate for the dearth of funding from traditional sources. Morgan Estabrook, *New Crowdfunding Site Allows Public to Advance U. Va. Research Projects Through Targeted Donations*, UVA TODAY (May 15, 2013), <http://news.virginia.edu/content/new-crowdfunding-site-allows-public-advance-uva-research-projects-through-targeted-donations>; Keith Brannon, *Tulane University Launches Crowdfunding Partnership for Medical Research*, TULANE U. (Dec. 10, 2013), [http://tulane.edu/news/releases/pr\\_12102013.cfm](http://tulane.edu/news/releases/pr_12102013.cfm). To enhance their likelihood of success and attract donors, those pursuing crowdfunding are well-advised to post convincing videos on funding websites and to follow up with blog entries and media coverage of their projects, to the extent possible. Perlstein, *supra*, at 561.

### C. TOOLS TO HELP PATIENTS NAVIGATE THE HEALTHCARE SYSTEM

Open health data can promote not only research but also services that are helpful for patients. Several enterprises are developing tools to help patients obtain suitable and affordable medical care. Aidin is a small startup that uses CMS data on health facilities and nursing homes to provide hospitals and patients with information about options for care after discharge from the hospital.<sup>116</sup> Aidin offers its clients listings of available providers, quality of care ratings, and reviews. It also helps hospitals track patient experiences and outcomes so that they can determine which providers are the best fit for patients with specific health conditions.<sup>117</sup>

Similarly, iTriage is a free mobile app and website that allows patients to look up their symptoms and learn about possible causes and treatments.<sup>118</sup> In addition, it assists patients in locating and selecting appropriate care options by providing a variety of information, including hospital wait times and physician ratings.<sup>119</sup> iTriage uses publicly available data from HHS, the FDA, and other sources.<sup>120</sup>

Other examples are the state all-payer claims databases, Medicare's Provider Utilization and Payment Data, and Medicare's Hospital Compare.<sup>121</sup> These educate patients about healthcare costs and quality and allow patients to compare prices for various inpatient and outpatient services.<sup>122</sup>

### D. GOVERNMENT TRANSPARENCY AND PUBLIC EDUCATION

Proponents of government transparency will be pleased by the proliferation of open data. Databases such as HealthData.gov, Genbank, and others<sup>123</sup> allow viewers to gain significant insight into the information that the government has collected about individuals and the healthcare

---

116. *Former Sec. Sebelius Celebrates Aidin in Annual "Health Datapalooza" Speech*, AIDIN (June 12, 2014), [http://www.myaidin.com/articles/june\\_2014/002.html](http://www.myaidin.com/articles/june_2014/002.html); *Our Story*, AIDIN, <http://www.myaidin.com/ourstory.html> (last visited Nov. 23, 2015).

117. *Our Story*, AIDIN, *supra* note 116.

118. *What is iTriage?*, ITRIAGE, <https://about.itriagehealth.com/itriage-what-is> (last visited Nov. 23, 2015).

119. *Id.*

120. *About Our Medical Content*, ITRIAGE, <https://about.itriagehealth.com/company-info/medical-content> (last visited Nov. 23, 2015).

121. *See supra* Section II.A.5; *Hospital Compare*, MEDICARE.GOV, <http://www.medicare.gov/hospitalcompare/search.html> (last visited Nov. 23, 2015).

122. *See Hospital Compare*, *supra* note 121.

123. *See supra* Part II.

industry. In some cases, such insight may generate public debate and critique of government investigative policies that could lead to positive policy changes.<sup>124</sup>

In addition, public-use data can function as an important educational tool.<sup>125</sup> Patients can research their own conditions, find doctors with special expertise, better prepare for their medical appointments, and assess different treatment options that they are given.<sup>126</sup> Furthermore, the general public can learn about the healthcare system, healthcare costs, disease trends, genetics, research and public health initiatives, and much more.<sup>127</sup> Ordinary citizens and students will be able to access raw data themselves and engage in research exercises, either within the framework of academic programs or on their own. For example, the New York University School of Medicine is leveraging open data resources to enhance its curriculum. It is creating patient snapshots from New York hospital discharge data and developing sophisticated training tools based on these cases.<sup>128</sup> Active learning and engagement with health data might also inspire greater public enthusiasm about medical research and more vocal support for government funding of this vital activity.

#### E. IMPROVEMENTS IN HEALTHCARE QUALITY AND PUBLIC HEALTH POLICY

Open data can fuel improvements in healthcare quality and health policies. A report from New York State provided a number of compelling illustrations.<sup>129</sup> In 2011, in preparation for Hurricane Irene, nursing home administrators used publicly available weekly bed census reports to identify facilities to which they could evacuate residents.<sup>130</sup> Likewise, annual

---

124. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 4 (stating that data sharing can “build trust with outside partners and the public by allowing open critique of CDC investigations”).

125. GRUSHKIN, *supra* note 6, at 4 (stating that “wider access to the tools of biotechnology, particularly those related to the reading and writing of DNA, has the potential to spur global innovation and promote biology education and literacy”).

126. Internet searches, however, should not replace consultation with medical experts, and often have pitfalls. Patients should not panic based on their independent research and become convinced that they suffer from a dreaded disease or have a poor prognosis before being examined by a physician. Patients also should not go to the doctor with a closed mind, unwilling to accept the expert’s own assessment and treatment recommendations.

127. *See supra* Part II.

128. Erika G. Martin et al., *Liberating Data to Transform Healthcare: New York’s Open Data Experience*, 311 JAMA 2481, 2481 (2014).

129. *Id.*

130. *Id.*

reports of cardiac surgery mortality rates, linked to the hospitals and surgeons who provide care, induced low-scoring facilities to undertake quality improvement initiatives and several physicians who performed poorly to leave practice.<sup>131</sup> A different study, published in 2015 in *Health Affairs*, concluded that Medicare's Hospital Compare "slowed the rate of price increases in a majority of states that had not previously been exposed to comparable information through their own public reporting systems."<sup>132</sup>

Once data are released, they are available not only to the general public, but also to the media. Media stories about health-related inequities can be particularly potent tools to effect policy changes. After officials released New York childhood obesity statistics that were organized by school district, news outlets highlighted the disparities in 2013 and some school administrators decided to improve their policies despite cost concerns.<sup>133</sup> A 2014 report in *Crain's New York Business* that publicized hospital cost disparities (for example, hip replacements that cost \$103,725 at New York University Hospitals Center but only \$15,436 at Bellevue Hospital Center) is likewise expected to catalyze pricing and reimbursement changes.<sup>134</sup>

#### IV. RISKS OF PUBLIC ACCESS TO HEALTH DATA

Although the benefits of opening health data resources to the public are considerable, the risks are not inconsequential. The federal research regulations cover only studies that are funded or conducted by federal government agencies or that do not use publicly available data.<sup>135</sup> Therefore, studies without federal funding and ones that use publicly available data are not subject to any formal oversight. Furthermore, the HIPAA Privacy Rule and state privacy laws most likely will not govern open databases.<sup>136</sup> This Part analyzes several potential risks associated with open access to patient-related health information: 1) privacy breaches; 2) discrimination and special targeting by employers, financial institutions, and marketers, among others; 3) propagation of incorrect and harmful research conclusions; and 4) litigation.

---

131. *Id.*

132. Avi Dor et al., *Medicare's Hospital Compare Quality Reports Appear to Have Slowed Price Increases for Two Major Procedures*, 34 HEALTH AFF. 71, 75 (2015) (focusing on coronary artery bypass grafts and percutaneous coronary interventions).

133. Martin et al., *supra* note 128, at 2481.

134. *Id.*

135. See 45 C.F.R. §§ 46.101(a), 46.101(b)(4) (2013).

136. See *infra* Section IV.A.1.

## A. PRIVACY THREATS

I recently logged onto the Personal Genome Project and looked at the Participant Profiles section.<sup>137</sup> To my surprise, several profiles disclosed the names of patients along with their date of birth, sex, weight, height, blood type, race, health conditions, medications, allergies, procedures, and more.<sup>138</sup> I wondered if these patients understood that anyone with a computer could view all of this information. Other profiles excluded the name of the participant but provided all of the other details, which could potentially allow a clever and motivated viewer to identify the patient.

Privacy threats are the first risk that may come to mind with respect to public use of patient-related medical big data. The HIPAA Privacy Rule,<sup>139</sup> the Privacy Act,<sup>140</sup> and numerous state privacy laws govern the disclosure of medical records.<sup>141</sup> However, the laws and regulations do not cover all data holders who make medical information publicly available.<sup>142</sup> In addition, public-use data is most often presented in de-identified form<sup>143</sup> and thus is exempt from the disclosure restrictions established in these laws and regulations.<sup>144</sup> Moreover, even with thorough de-identification, at least a small risk of re-identification remains. Privacy concerns thus deserve thorough analysis.

### 1. *Privacy Law*

Many federal and state laws address medical privacy. None of the laws, however, provide patients with comprehensive protection, and even in the aggregate, they leave many gaps. The following discussion describes laws and regulations relevant to the disclosure of patient-related data for public use.

---

137. *Participant Profiles*, PERSONAL GENOME PROJECT, HARV. MED. SCH., <https://my.pgp-hms.org/users> (last visited Nov. 23, 2015).

138. *Public Profile -- hu43860C*, PERSONAL GENOME PROJECT, HARV. MED. SCH., <https://my.pgp-hms.org/profile/hu43860C> (last updated Sept. 4, 2015).

139. 45 C.F.R. §§ 160.101–.534 (2013).

140. 5 U.S.C. § 552a (2010).

141. *See* AMERICANS HEALTH LAWYERS ASSOCIATION, STATE HEALTHCARE PRIVACY LAW SURVEY (2013); Sarah Hexem, *Public Health Departments and State Patient Confidentiality Laws Map*, LAWATLAS, <http://lawatlas.org/preview?dataset=public-health-departments-and-state-patient-confidentiality-laws> (last visited Nov. 23, 2015).

142. *See infra* Sections IV.A.1 and IV.A.3.a.

143. *See supra* Part II.

144. *See infra* Section IV.A.1.

## a) The HIPAA Privacy Rule

The HIPAA Privacy Rule establishes that, with some exceptions, entities covered by the regulations must obtain patients' permission before disclosing their medical information to third parties.<sup>145</sup> The Rule, however, covers only health plans, healthcare clearinghouses, healthcare providers who transmit health information electronically for purposes of HIPAA-relevant transactions, and their business associates.<sup>146</sup> It does not apply to government agencies or private enterprises that are not acting in these capacities. Thus, HIPAA does not regulate many of the websites discussed in Part II of this Article, such as those operated by state governments, CDC, Dryad or PatientsLikeMe.

Moreover, the HIPAA Privacy Rule protects only "individually identifiable health information" that is electronically or otherwise transmitted or maintained.<sup>147</sup> Consequently, the federal regulations do not govern data that custodians de-identify<sup>148</sup> and open to the public.

## b) The Privacy Act

The Privacy Act is a federal law that governs the collection, storage, use, and disclosure of information by the federal government.<sup>149</sup> The law provides that the federal government may not disclose records without the data subject's permission, unless specific exceptions apply. However, the Privacy Act defines the term "record" as an item that contains a person's "name, or the identifying number, symbol, or other identifying particular assigned to the individual."<sup>150</sup> Consequently, the Privacy Act exempts the government's dissemination of de-identified information on HealthData.gov or other websites.

## c) State Laws

All states have recognized a common law or statutory right to privacy<sup>151</sup> and have statutes that address privacy concerns.<sup>152</sup> A thorough

---

145. 45 C.F.R. §§ 164.508–.510 (2013).

146. 45 C.F.R. §§ 160.102–.103 (2013); 42 U.S.C. § 17934 (2010).

147. 45 C.F.R. § 160.103 (2013).

148. See *infra* Section IV.A.2 (discussing HIPAA's requirements for de-identification).

149. 5 U.S.C. § 552a (2010).

150. *Id.* at § 552a(a)(4).

151. See Corrine Parver, *Patient-Tailored Medicine, Part Two: Personalized Medicine and the Legal Landscape*, 2 J. HEALTH & LIFE SCI. L. 1, 32 (2009).

152. See AMERICANS HEALTH LAWYERS ASSOCIATION, *supra* note 141; LAWATLAS, *supra* note 141.

analysis of state law is beyond the scope of this Article.<sup>153</sup> In general, state laws are varied and inconsistent, often providing piecemeal protection for some types of data but not others.<sup>154</sup> Moreover, like the HIPAA Privacy Rule and the Privacy Act, states typically allow disclosure of de-identified health information without patient authorization.<sup>155</sup> Therefore, most of the public-use data resources contemplated in this Article would not be governed by state law.

## 2. *De-identification*

The foregoing discussion raises the following critical question: what does “de-identified” mean, and how can data holders achieve de-identification? The HIPAA Privacy Rule provides a detailed answer. It states that health information is de-identified if (1) a qualified expert determines that there is only a “very small” risk that the data can be re-identified, and (2) the expert documents his or her analysis.<sup>156</sup> The Department of Health and Human Services issued guidance that endorsed several de-identification techniques:

- *Suppression*, which involves redaction of particular data features prior to disclosure (e.g., removing zip codes, birthdates, income);
- *Generalization*, which involves transforming particular information into less specific representations (e.g., indicating a 10-year age range instead of exact age); and
- *Perturbation*, which involves exchanging certain data values for equally specific but different values (e.g., changing patients’ ages).<sup>157</sup>

---

153. For detailed information about state privacy and confidentiality laws, see AMERICANS HEALTH LAWYERS ASSOCIATION, *supra* note 141; LAWATLAS, *supra* note 141.

154. See Deven McGraw et al., *Privacy as an Enabler, Not an Impediment: Building Trust into Health Information Exchange*, 28 HEALTH AFF. 416, 420 (2009) (noting that “[a]lthough the states have an important role to play in privacy policy, state privacy laws are fragmentary and inconsistent, providing neither developers nor consumers with the assurances they deserve, especially for services of nationwide reach”).

155. Scott Burris et al., *The Role of State Law in Protecting Human Subjects of Public Health Research and Practice*, 31 J.L. MED. & ETHICS 654, 656 (2003).

156. 45 C.F.R. § 164.514(b)(1) (2013).

157. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, U.S. DEPT’ OF HEALTH & HUMAN SERVS. (Nov. 26, 2012), <http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html>

In the alternative, according to the HIPAA Privacy Rule, de-identification is achieved if the following eighteen identifiers are removed:

- (A) Names;
  - (B) All geographic subdivisions smaller than a State, including street address, city, county, precinct, zip code, and their equivalent geocodes, except for the initial three digits of a zip code if, according to the current publicly available data from the Bureau of the Census:
    - (1) The geographic unit formed by combining all zip codes with the same three initial digits contains more than 20,000 people; and
    - (2) The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people is changed to 000;
  - (C) All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older;
  - (D) Telephone numbers;
  - (E) Fax numbers;
  - (F) Electronic mail addresses;
  - (G) Social security numbers;
  - (H) Medical record numbers;
  - (I) Health plan beneficiary numbers;
  - (J) Account numbers;
  - (K) Certificate/license numbers;
  - (L) Vehicle identifiers and serial numbers, including license plate numbers;
  - (M) Device identifiers and serial numbers;
  - (N) Web Universal Resource Locators (URLs);
  - (O) Internet Protocol (IP) address numbers;
  - (P) Biometric identifiers, including finger and voice prints;
  - (Q) Full face photographic images and any comparable images;
- and

---

#guidancedetermination (noting that techniques such as suppression and generalization are often used in combination).

- (R) Any other unique identifying number, characteristic, or code . . . .<sup>158</sup>

Health information that has all eighteen identifiers removed in accordance with the HIPAA “safe harbor” provision is considered per se de-identified and exempted from HIPAA coverage unless a covered entity knows that the data can be used on its own or together with other information to identify a data subject.<sup>159</sup> For example, if researchers request only data pertaining to a very small geographic area in which most people know each other, it may be impossible to truly de-identify the information.<sup>160</sup> In such a case, experts may need to aggregate data from several locations or to combine suppression with other techniques.

### 3. *Does Public-Use Medical Data Pose a Real Privacy Threat?*

Data custodians offering public-use data may try hard to de-identify patient records or to ask for patients’ consent to disclosure.<sup>161</sup> Nevertheless, many are not required to do so because they are not covered by the HIPAA Privacy Rule and its data disclosure and de-identification guidelines. Consequently, the patient authorization and de-identification practices that data custodians choose to implement may deviate from HIPAA standards and leave data more vulnerable to attack by hackers or other wrongdoers.

Moreover, even with careful de-identification, sophisticated adversaries may be able to re-identify at least a small number of records. Successful de-identification of genetic information may be particularly challenging. With voluminous de-identified medical data available to the public, re-identification attempts are likely to occur. Perpetrators may have malevolent intent, such as identity theft, or may simply be interested in determining whether they can meet the challenge of re-identification.

#### a) Data Holders Not Covered by the HIPAA Privacy Rule

The HIPAA Privacy Rule’s health information disclosure and de-identification requirements do not apply to most suppliers of publicly available health data, because they are either government agencies or non-

---

158. 45 C.F.R. § 164.514(b)(2)(i) (2013). Removal of the eighteen identifiers is a comprehensive form of suppression.

159. 45 C.F.R. § 164.514(b)(2)(ii) (2013).

160. Khaled El Emam et al., *Evaluating Predictors of Geographic Area Population Size Cut-offs to Manage Re-identification Risk*, 16 J. AM. MED. INFORMATICS ASS’N 256, 256–57 (2009); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1156 (2013).

161. *See supra* Part II.

covered private entities.<sup>162</sup> Consequently, these data holders may not be diligent about obtaining meaningful patient authorization for disclosure of identifiable information. In addition, if they de-identify records, they may choose to do so in ways that provide far less privacy protection to their subjects than does the HIPAA safe harbor provision. Stripping medical records of names alone does little to conceal patients' identities, and even leaving just a few specific details may make it easy to ascertain who the individual is. One startling study found that almost 98% of Montreal residents could be identified based on their full postal code, date of birth, and gender.<sup>163</sup>

Data holders' de-identification practices vary. A 2013 survey found that thirty-three states released patient hospital discharge data to the public, but only seven de-identified them in a manner that would conform to the HIPAA Privacy Rule's standard.<sup>164</sup> Many states released the month or quarter of hospital admission and/or discharge and patients' five-digit zip codes.<sup>165</sup> Datasets with these details are more vulnerable to re-identification than those that are de-identified in accordance with HIPAA guidance. The more personal details a publicly available health record contains, the more likely it is to be matched to other open datasets that include names, such as voter registration lists, purchasing records,<sup>166</sup> or news reports.<sup>167</sup> Thus, the more overlapping information fields there are between the medical records and other datasets, such as zip codes, ages, and details of illness, the more likely an adversary will be able to link names to the purportedly anonymized health information.

Scholars confirm that concern about re-identification is well-grounded, as demonstrated by a variety of re-identification successes. In a particularly infamous case, Latanya Sweeney, now a computer scientist at Harvard University, identified the health records of Massachusetts' Governor William Weld when she was a graduate student at the

---

162. *See supra* note 146 and accompanying text.

163. Khaled El Emam, *The Re-identification Risk of Canadians from Longitudinal Demographics*, 11 BMC MED. INFORMATICS & DECISION MAKING 46, 51 (2011).

164. SEAN HOOLEY & LATANYA SWEENEY, SURVEY OF PUBLICLY AVAILABLE STATE HEALTH DATABASES 4 (2013), <http://dataprivacylab.org/projects/50states/1075-1.pdf>.

165. *Id.* at 4–7.

166. *See infra* note 196 and accompanying text (discussing information that third parties can purchase about individuals).

167. Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of "Personally Identifiable Information,"* COMM. ACM, June 2010, at 24, 26; *Re-identification*, ELECTRONIC PRIVACY INFORMATION CENTER, <http://epic.org/privacy/reidentification> (last visited Nov. 23, 2015).

Massachusetts Institute of Technology in 1996.<sup>168</sup> She compared birth date, gender, and zip code information that was retained in publicly released hospital discharge records to the same identifiers in publicly available voter registration lists and could match voter names to hospital records.<sup>169</sup>

In a more recent effort, Dr. Sweeney and colleagues worked to re-identify publicly available profiles in the Personal Genome Project<sup>170</sup> that contained medical and genomic information as well as date of birth, gender, and zip code.<sup>171</sup> They linked the demographic data to voter lists or other public records that featured names and were able to identify eighty-four to ninety-seven percent of Personal Genome Project profiles.<sup>172</sup>

In a third project, Dr. Sweeney focused on Washington State hospital discharge data, which contained many demographic details other than names and addresses and could be purchased for fifty dollars. She attempted to match hospitalization records to eighty-one newspaper stories about accidents and injuries in 2011 and was able to determine the name of the patient to whom the records belonged in thirty-five (or forty-three percent) of the cases, based on the news accounts.<sup>173</sup>

#### b) Re-identification of Fully De-identified Health Records

Theoretically, de-identification in accordance with the HIPAA Privacy Rule's guidelines should make it impossible for anyone to determine the identity of data subjects. Nevertheless, experts have concluded that there remains a small risk that highly skilled and motivated attackers will be able to re-identify records that have been de-identified in

---

168. Jonathan Shaw, *Exposed: The Erosion of Privacy in the Internet Era*, HARV. MAG., Sept.–Oct. 2009, at 38, <http://harvardmagazine.com/2009/09/privacy-erosion-in-internet-era>.

169. *Id.*; Kathleen Benitez & Bradley Malin, *Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule*, 17 J. AM MED. INFORMATICS ASS'N 169, 169 (2010).

170. *See supra* Section II.B.3.

171. Latanya Sweeney et al., *Identifying Participants in the Personal Genome Project by Name* (Harv. U. Data Privacy Lab, White Paper 1021-1, Apr. 24, 2013), <http://dataprivacylab.org/projects/pgp/1021-1.pdf>.

172. *Id.* at 1. The researchers found that some Personal Genome Project profiles contained the data subject's name, and in other instances, when the downloadable DNA files were uncompressed, they had a file name that included the data subjects' first and last names. *Id.* at 3.

173. Latanya Sweeney, *Matching Known Patients to Health Records in Washington State Data* (Harv. U. Data Privacy Lab, White Paper 1089-1, July 4, 2013), <http://dataprivacylab.org/projects/wa/1089-1.pdf>; Jordan Robertson, *States' Hospital Data for Sale Puts Privacy in Jeopardy*, BLOOMBERG (Jun 4, 2013), <http://www.bloomberg.com/news/2013-06-05/states-hospital-data-for-sale-puts-privacy-in-jeopardy.html>.

compliance with HIPAA guidelines.<sup>174</sup> Re-identification may occur when perpetrators have access to non-medical open data, such as voter registration records, that they can link to anonymized health information. Studies have estimated that the risk of re-identification of HIPAA de-identified records falls in the range of 0.01% to 0.25%.<sup>175</sup> Although this percentage seems tiny, it translates into a risk of tens of thousands or even hundreds of thousands of records being re-identified if one thinks in terms of the 323 million individuals in the American population.<sup>176</sup>

Furthermore, the HIPAA Privacy Rule's safe harbor provision does not ban the disclosure of certain details whose presence could make it easier to identify individuals. For example, according to Dr. Khaled El Emam, if hospital discharge data includes length of stay and time since last visit, which are not among the eighteen prohibited identifiers, as many as 16.57% of the records could have a high likelihood of re-identification.<sup>177</sup>

### c) The Peculiarities of Genetic Information

The HIPAA Privacy Rule does not provide explicit guidance concerning the de-identification of genetic information,<sup>178</sup> such as the genetic sequences available through GenBank.<sup>179</sup> Many commentators have expressed concern that adversaries could re-identify anonymized genetic information using a variety of techniques.<sup>180</sup> Researchers believe that people can be uniquely identified through a sequence of only thirty to eighty out of thirty million single-nucleotide polymorphisms (SNPs).<sup>181</sup> In

---

174. Hoffman & Podgurski, *supra* note 15, at 105–07.

175. Khaled El Emam et al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLOS ONE e28071 (2011), <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028071> (finding a re-identification rate of 0.013%); NAT'L COMM. ON VITAL & HEALTH STATISTICS, ENHANCED PROTECTIONS FOR USES OF HEALTH DATA 36 n.16 (2007), <http://www.ncvhs.hhs.gov/wp-content/uploads/2014/05/071221t.pdf>.

176. See *U.S. and World Population Clock*, U.S. CENSUS BUREAU, <http://www.census.gov/popclock> (last visited Feb. 5, 2016).

177. Khaled El Emam, *Methods for the De-identification of Electronic Health Records for Genomic Research*, 3 GENOME MED. 25, 27 (2011).

178. 45 C.F.R. § 164.514(b)(2)(i) (2013); El Emam, *supra* note 177, at 27.

179. See *supra* Section II.A.4; Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321, 321 (2013) (noting that “[s]haring sequencing data sets without identifiers has become a common practice in genomics”).

180. El Emam, *supra* note 177, at 27; Dina N. Paltoo et al., *Data Use Under the NIH GWAS Data Sharing Policy and Future Directions*, 46 NATURE GENETICS 934, 937 (2014).

181. El Emam, *supra* note 177, at 27; Liina Kamm et al., *A New Way to Protect Privacy in Large-Scale Genome-Wide Association Studies*, 29 BIOINFORMATICS 886, 886

one study, researchers identified family names by matching short sequences of DNA bases on an individual's Y chromosome to entries in recreational genetic genealogy databases.<sup>182</sup> These short sequences are repeated different numbers of times in different individuals, and hence they are called short tandem repeats or Y-STRs. Even providing only summary-level genetic information cannot always fully protect the identities of data subjects.<sup>183</sup> Given genotype frequencies for a study cohort, it is possible to determine if a particular individual is in the cohort if one knows the individual's genotype and has a reference set of allele frequencies for the underlying population.<sup>184</sup> Thus, genetic information may be more difficult to de-identify effectively than other types of data.

#### B. DISCRIMINATION AND SPECIAL TARGETING

Medical big data can serve as a treasure trove of information for parties who will use it to further their own economic interests.<sup>185</sup> The release of patient data for public use, alongside advances in re-identification capabilities, raises significant concern regarding potential discrimination or targeting by parties with a stake in individuals' health and economic welfare.<sup>186</sup> This Section will focus on three examples: employers, financial institutions, and marketers. Employers have a strong incentive to identify

---

(2013). A single-nucleotide polymorphism is a "variation at a single position in a DNA sequence among individuals." *Single Nucleotide Polymorphism*, NATURE, <http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295> (last visited Nov. 23, 2015).

182. Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321, 321 (2013).

183. David W. Craig et al., *Assessing and Managing Risk When Sharing Aggregate Genetic Variant Data*, 12 NATURE REV. GENETICS 730, 730 (2012).

184. *Id.* at 734–35. An allele is one of several variations of a gene. *Allele*, GENETICS HOME REFERENCE, U.S. NAT'L LIBR. OF MED. (Feb. 1, 2016), <http://ghr.nlm.nih.gov/glossary=allele>.

185. See Narayanan & Shmatikov, *supra* note 167, at 26 (noting "increasing economic incentives for potential attackers"); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 96–99 (2014) (discussing business use of big data to obtain personal health information about consumers); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 3 (2014) (stating that in today's world "[p]redictive algorithms mine personal information to make guesses about individuals' likely actions and risks[]" and "[p]rivate and public entities rely on predictive algorithmic assessments to make important decisions about individuals").

186. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51 (2014), [http://www.whitehouse.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf) (stating that "[a]n important conclusion of this study is that big data technologies can cause societal harms beyond damages to privacy, such as discrimination against individuals and groups").

and select the healthiest workers in order to avoid attendance and productivity problems and high health insurance costs. Likewise, lenders are interested in borrowers who will have income and be able to pay off their loans. For their part, advertisers and marketers wish to tailor their marketing campaigns to reach the most lucrative markets, and thus, they might target particular individuals based on known health conditions<sup>187</sup> or offer special promotions to some consumers but not others.<sup>188</sup>

### 1. *Employers*

Employers go to great lengths to select employees carefully in order to maximize business productivity and profitability. Sick or disabled employees can be very expensive for employers because of absenteeism, performance shortcomings, high insurance costs, loss of customers who are uncomfortable interacting with the individual, erosion of workforce morale if other workers feel overburdened while the employer accommodates the ill or impaired employee, and other problems.<sup>189</sup> Employers may have good economic reasons to strive for the healthiest possible workforce, but they are constrained by federal and state laws prohibiting discrimination based on a variety of protected classifications, including disability and genetic information.<sup>190</sup> Moreover, if employers make assumptions about people's health and apply rigid, generalized rules to determine which employees are undesirable, they will deprive many qualified individuals of job opportunities.

The advent of publicly available data may enable employers to discriminate against individuals who are perceived to be at high risk of poor health in ways that are subtle and difficult to detect. Some employers are already embracing advanced technologies such as smart badges that enable them to monitor employee conduct and analyze workplace

---

187. Lori Andrews, *Facebook is Using You*, N.Y. TIMES (Feb. 4, 2012), <http://www.nytimes.com/2012/02/05/opinion/sunday/facebook-is-using-you.html>.

188. EXEC. OFFICE OF THE PRESIDENT, BIG DATA, *supra* note 186, at 47.

189. See Bruce Japsen, *U.S. Workforce Illness Costs \$576B Annually From Sick Days To Workers Compensation*, FORBES (Sept. 12, 2012), <http://www.forbes.com/sites/brucejapsen/2012/09/12/u-s-workforce-illness-costs-576b-annually-from-sick-days-to-workers-compensation>; Jessica L. Roberts, *Healthism and the Law of Employment Discrimination*, 99 IOWA L. REV. 571, 580–89 (2014) (analyzing the rationales for health-driven employment policies).

190. See Sharona Hoffman, *The Importance of Immutability in Employment Discrimination Law*, 52 WM. & MARY L. REV. 1483, 1489–94 (2011) (discussing the forms of discrimination prohibited by anti-discrimination legislation).

interactions as never before.<sup>191</sup> They may well pursue opportunities to use identifiable, re-identifiable, and even non-identifiable medical data to develop new screening tools and hiring policies.

a) Using Identifiable or Re-Identifiable Data

Individuals who agree to share identifiable or easily re-identifiable medical data with the public on websites such as PatientsLikeMe or the Personal Genome Project<sup>192</sup> should understand that it will be accessible to everyone. This includes not only fellow patients or others with benign interests, but also employers who may take adverse action based on health concerns.

Many employers reportedly access public profiles that applicants post on social media sites as part of their investigation of candidates' credentials.<sup>193</sup> They also ask applicants for permission to obtain their credit reports.<sup>194</sup> It is therefore not far-fetched to assume that employers will search publicly available health profiles as well. It is also possible that employers will hire data miners to re-identify medical information when doing so is not excessively difficult. Employers or their agents may be able to re-identify health records that feature certain items such as postal codes, birthdates, and gender, with the aid of demographic information and names contained in voter registration lists, credit reports, or job applications.<sup>195</sup>

Employers may also be able to hire experts who can re-identify information that is thoroughly de-identified in compliance with the HIPAA safe harbor standard if they have a sufficient amount of related, identifiable data about applicants and employees to which they can match the de-identified records. For example, data miners may be able to obtain individuals' detailed purchasing histories or web-browsing histories from

---

191. Steve Lohr, *Unblinking Eyes Track Employees: Workplace Surveillance Sees Good and Bad*, N.Y. TIMES (June 21, 2014), <http://www.nytimes.com/2014/06/22/technology/workplace-surveillance-sees-good-and-bad.html>.

192. *See supra* Sections II.B.2 and II.B.3.

193. Greg Fish & Timothy B. Lee, *Employer Get Outta My Facebook*, BLOOMBERG BUSINESSWEEK (Mar. 20, 2008), [http://www.businessweek.com/debateroom/archives/2010/12/employers\\_get\\_outta\\_my\\_facebook.html](http://www.businessweek.com/debateroom/archives/2010/12/employers_get_outta_my_facebook.html); Phyllis Korkki, *Is Your Online Identity Spoiling Your Chances?*, N.Y. TIMES (Oct. 9, 2010), <http://www.nytimes.com/2010/10/10/jobs/10search.html>.

194. Gary Rivlin, *The Long Shadow of Bad Credit*, N.Y. TIMES (May 12, 2013), <http://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html>.

195. *See supra* Section IV.A.3.a.

database marketers such as Acxiom,<sup>196</sup> and by some estimates, approximately 4000 data brokers already exist.<sup>197</sup> If these lists suggest that particular workers have certain health conditions, data miners may be able to link anonymized health records to names on the lists and thereby identify patients and obtain their medical details.

Experienced data miners, aided by contemporary technology, often have no difficulty achieving re-identification. Interested buyers can purchase lists of patients with depression, erectile dysfunction, diabetes, Alzheimer's disease, and Parkinson's disease.<sup>198</sup> In a 2010 article, two computer scientists, Arvind Narayanan and Vitaly Shmatikov, went as far as to say that "advances in the art and science of re-identification, increasing economic incentives for potential attackers, and ready availability of personal information about millions of people (for example, in online social networks) are rapidly rendering [de-identification] obsolete."<sup>199</sup>

The Americans with Disabilities Act (ADA) prohibits employers from engaging in disability-based discrimination.<sup>200</sup> The law allows employers to conduct medical inquiries and examinations within certain limits to determine fitness for duty,<sup>201</sup> but workers who feel that an employer denied them a job opportunity because of information it discovered may sue the employer.<sup>202</sup> Unlike medical exams, publicly shared medical data would enable employers to view workers' health information without the individuals' knowledge and, consequently, with little concern about being

---

196. See Alice E. Marwick, *How Your Data Are Being Deeply Mined*, N.Y. REV. BOOKS, Jan. 9, 2014, <http://www.nybooks.com/articles/archives/2014/jan/09/how-your-data-are-being-deeply-mined> (discussing the development of "database marketing," an industry that collects, aggregates, and brokers personal data from sources such as "home valuation and vehicle ownership, information about online behavior tracked through cookies, browser advertising, and the like, data from customer surveys, and 'offline' buying behavior"); see also ACXIOM, <http://www.acxiom.com> (last visited Nov. 23, 2015) (describing a company that gives "clients the power to successfully manage audiences, personalize customer experiences and create profitable customer relationships" using big data analytics).

197. Frank Pasquale, *The Dark Market for Personal Data*, N.Y. TIMES (Oct. 17, 2014), <http://www.nytimes.com/2014/10/17/opinion/the-dark-market-for-personal-data.html>.

198. Shannon Pettypiece & Jordan Robertson, *For Sale: Your Name and Medical Condition*, BLOOMBERG BUSINESS (Sept. 18, 2014), <http://www.bloomberg.com/bw/articles/2014-09-18/for-sale-your-name-and-medical-condition>.

199. Narayanan & Shmatikov, *supra* note 167, at 26; see also ELECTRONIC PRIVACY INFORMATION CENTER, *supra* note 167 (stating that "'anonymized' data can easily be re-identified").

200. 42 U.S.C. § 12112(a) (2010).

201. 42 U.S.C. § 12112(d) (2010).

202. 42 U.S.C. § 12117(a) (2010).

accused of disability discrimination in case of adverse employment decisions.

b) De-identified Information as a Basis for Multi-Factor  
Discrimination and Discrimination by Proxy

Employers may use publicly available medical data for purposes of screening workers even without attempting to re-identify records. Some websites feature information concerning disease trends that might induce employers to try to exclude certain classes of employees. For instance, CDC Wonder allows users to search for cancer incidence by age, sex, race, ethnicity, and region.<sup>203</sup> As a hypothetical example, the results of a search might lead an employer to conclude that Hispanic women over fifty are more prone to several cancers than other individuals, and consequently, to decline to hire Hispanic women over fifty.<sup>204</sup>

Some researchers have in fact focused on particular ethnic sub-groups and concluded that they have more health problems than others. A prime example is the PINE Study, for which investigators interviewed 3,018 Chinese adults aged 60 to 105 who lived in the Chicago area between 2011 and 2013.<sup>205</sup> The study concluded that “Chinese older adults experience disproportionate health disparities,” suffering from significant physical, psychological, financial, and social challenges.<sup>206</sup> Though this was far from the study’s intention, readers of the report may think twice about hiring people of Chinese ancestry who are sixty or older. While investigators used interviews for this study, they could also undertake record reviews in the future if sufficient information is available. The study’s findings could encourage employers to pursue similar research using open medical data, because it will yield clear categories of individuals who should be excluded as likely to become problematic employees.

The civil rights laws prohibit discrimination by race, color sex, and age, among other categories,<sup>207</sup> but discrimination based on a combination of two or more factors would be very difficult to detect and prove. If accused of discrimination, the employer would be able to show that it has

---

203. *United States Cancer Statistics, 1999–2010 Incidence Archive Request*, CDC WONDER, <http://wonder.cdc.gov/cancer-v2010.html> (last visited Nov. 23, 2015).

204. See Jourdan Day, *Closing the Loophole—Why Intersectional Claims are Needed to Address Discrimination Against Older Women*, 75 OHIO ST. L.J. 447, 448 (2014).

205. XINQI DONG ET AL., *THE PINE REPORT*, at v (2013), [http://chinesehealthyaging.org/files/PINE\\_Final\\_Reports/All.pdf](http://chinesehealthyaging.org/files/PINE_Final_Reports/All.pdf).

206. *Id.* at v, 40.

207. See Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a) (2006); Age Discrimination in Employment Act, 29 U.S.C. §§ 623(a), 631(a) (2006).

Hispanic, female, and older employees in its workforce. A plaintiff would need to be clever enough to discern that the employer is excluding only a subgroup that falls at the intersection of several protected categories and then somehow decipher the employer's motivation for doing so. Furthermore, many courts disallow multi-factor claims involving age.<sup>208</sup> These courts perceive "age plus" cases as prohibited by a Supreme Court decision, *Gross v. FBL Financial Services, Inc.*, that held that a plaintiff claiming age discrimination must prove that age was the "but for" reason for the adverse action at issue.<sup>209</sup>

Anonymized data can provide other opportunities for discrimination as well.<sup>210</sup> Employers, who are highly motivated to develop means to screen out workers at high risk of health problems, may undertake their own citizen science projects or hire experts to do so. Employers or their agents may mine medical data using sophisticated algorithms to detect associations between individual characteristics or behaviors and poor physical or mental health.<sup>211</sup> Then, through job applications, interviews, and reference or background checks, employers could try to determine whether applicants have those attributes or behaviors.

Concern that employers would attempt to find reliable predictors of applicants' future health status is not fanciful. In the words of two prominent scholars, "predictive algorithms . . . are increasingly rating people in countless aspects of their lives."<sup>212</sup> Several websites, such as "Lifespan Calculator" and "How Long Will I Live?," invite users to calculate their longevity based on a series of questions. These websites' calculations may or may not be trustworthy or illuminating, but they reflect deep interest in creating health-related predictive tools.<sup>213</sup> The websites ask users about their height, weight, education, income, marital status, exercise habits, smoking, drinking, driving, seat belt use, work history, eating, sleeping, and more.<sup>214</sup> They also ask a small number of

---

208. Day, *supra* note 204, at 449.

209. *Id.* at 466–67; *Gross v. FBL Fin. Servs., Inc.*, 557 U.S. 167, 177–78 (2009).

210. Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HARV. BUS. REV. (Jan. 29, 2014), <http://blogs.hbr.org/2014/01/big-datas-dangerous-new-era-of-discrimination>.

211. See EXEC. OFFICE OF THE PRESIDENT, BIG DATA, *supra* note 186, at 45–47 (discussing algorithms).

212. Citron & Pasquale, *supra* note 185, at 2.

213. See *Lifespan Calculator*, NORTHWESTERN MUTUAL, <http://media.nmfn.com/network/lifespan> (last visited Nov. 23, 2015); Dean P. Foster, Choong Tze Chua, & Lyle Y. Ungar, *How Long Will I Live?*, U. PENN., <http://gosset.wharton.upenn.edu/mortality/perl/CalcForm.html> (last visited Nov. 23, 2015).

214. See *Lifespan Calculator*, *supra* note 213; *How Long Will I Live?*, *supra* note 213.

questions about family and personal medical history. If employers asked such questions directly, they could be found liable for violations of federal anti-discrimination law.<sup>215</sup> However, as data mining science continues to develop and demand for its products grows, experts will likely develop dependable tools that do not require such explicit questions. While employers may not care about whether employees will live to be eighty or ninety, they will be interested in determining whether they will remain healthy and productive during their working lives.

Already, some employers are known to reject candidates who are obese or smoke because of anticipated health problems.<sup>216</sup> In the future, they might disqualify applicants for many more forms of conduct or characteristics. Applicants could routinely be questioned during interviews about their eating, exercise, travel, and other habits. Employers may then base employment decisions on proxies for disease or predictions of later illness without violating state and federal anti-discrimination laws. As Professor Jessica Roberts explains, those statutes prohibit discrimination based on attributes (for example, race or disability) rather than on behavior (for example, consumption of fatty food or a sedentary lifestyle).<sup>217</sup> Furthermore, the laws focus only on *current* disabilities and genetic information and do not govern any assumptions employers might make about individuals' future ailments that do not relate to off-limits genetic information.<sup>218</sup>

## 2. *Financial Institutions and Marketers*

Like employers, financial institutions collect information about individuals. Banks routinely maintain databases with data about customers who previously overdrew their accounts or bounced checks.<sup>219</sup> Nothing will

---

215. See Genetic Information Nondiscrimination Act, 42 U.S.C. § 2000ff(4) (2008) (including “the manifestation of a disease or disorder in family members” in the definition of “genetic information” that employers are forbidden to seek); Americans with Disabilities Act, 42 U.S.C. § 12112(d)(2) (prohibiting employers from conducting most medical inquiries and tests prior to extending a job offer to the applicant).

216. Roberts, *supra* note 189, at 577–79.

217. *Id.* at 604–07.

218. See Hoffman, *supra* note 190, at 1489–94 (2011) (discussing the forms of discrimination prohibited by anti-discrimination legislation). The Genetic Information Nondiscrimination Act prohibits employers from discriminating based on genetic information, and therefore, employers should refrain from mining data collections for genetic information, even if it is abundantly available. Genetic Information Nondiscrimination Act, 42 U.S.C. §§ 2000ff(4), 2000ff-1(a) (2008).

219. Jessica Silver-Greenberg & Michael Corkery, *Bank Account Screening Tool is Scrutinized as Excessive*, N.Y. TIMES (June 15, 2014), <http://dealbook.nytimes.com/2014/06/15/bank-account-screening-tool-is-scrutinized-as-excessive>.

prevent them from adding health information to their databases in order to hone their ability to screen out applicants with a high risk of defaulting on loans if such data is attainable at low cost. As suggested above, financial institutions may utilize identifiable and easily re-identifiable information and may mine databases to discern associations between health risks and various attributes or behaviors.<sup>220</sup>

The ADA prohibits disability-based discrimination by places of public accommodation, that is, establishments that provide services to the public, including banks and other financial institutions.<sup>221</sup> However, customers are unlikely to suspect or discover that banks viewed their health information while assessing their loan applications and thus, such acts of discrimination will probably go unchallenged.

Marketers and advertisers also have an interest in individuals' health data. The more they know about potential customers, the more they can tailor their materials to appeal to those individuals.<sup>222</sup> For example, individuals who are known to have diabetes might receive advertisements about sugar-free products, which some may perceive as a troubling invasion of privacy. Consumers may be particularly resentful when the health condition at issue is sensitive, as noted in a 2012 *Forbes* magazine article entitled "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did."<sup>223</sup>

Marketers may also engage in discriminatory practices, offering promotions and discounts to some customers but not others, or advertising selectively so that they reach only certain consumers. They may mine health records for clues regarding individuals' purchasing potential and aggressively pursue the most likely or wealthiest customers. A 2014 presidential report provided the following account:

[S]ome . . . retailers were found to be using an algorithm that generated different discounts for the same product to people based on where they believed the customer was located. While it may be that the price differences were driven by the lack of competition in certain neighborhoods, in practice, people in

---

220. See *supra* Section IV.B.1 (discussing potential discrimination by employers).

221. 42 U.S.C. §§ 12181(7)(F), 12182(a) (2010).

222. Andrews, *supra* note 187.

223. Kashmir Hill, *How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did*, *FORBES* (Feb. 16, 2012), <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did> (discussing Target's practice of data-mining its customers' purchasing records in order "to figure out what you like, what you need, and which coupons are most likely to make you happy").

higher-income areas received higher discounts than people in lower-income areas.<sup>224</sup>

While this practice already exists, access to open medical data may enable industry to refine marketing campaigns even further, to the dismay of some customers. Moreover, courts are unlikely to find that selective advertising or promotional offers and discounts violate anti-discrimination laws.<sup>225</sup> Marketers will generally be able to argue convincingly that their decisions were based on economic factors rather than on race, disability, or other protected categories.<sup>226</sup>

### C. PROPAGATION OF INCORRECT AND HARMFUL RESEARCH CONCLUSIONS

Citizen science can lead to valuable and illuminating discoveries.<sup>227</sup> At the same time, however, amateurs may reach incorrect conclusions.<sup>228</sup> Furthermore, anyone can widely publicize information on the Internet, whether it be correct or erroneous. Advice as to how to gain broad exposure is abundantly available on the Internet and can be found in webpages such as “12 Ways to Promote Your Blog”<sup>229</sup> and “How to Promote Your Article Online.”<sup>230</sup> In some cases, the media, celebrities, and politicians highlight the work of ordinary citizens,<sup>231</sup> and they may well do so with respect to scientific discoveries that they find intriguing or that support their own agendas. In other cases, individuals can gain

224. EXEC. OFFICE OF THE PRESIDENT, BIG DATA, *supra* note 186, at 46–47.

225. Schrage, *supra* note 210 (stating that it is unclear “where value-added personalization and segmentation end and harmful discrimination begins”).

226. Crawford & Schultz, *supra* note 185, at 101 (stating that “housing providers could design an algorithm to predict the relevant PII [personally identifiable information] of potential buyers or renters and advertise the properties only to those who fit these profiles” and do so without violating fair housing laws).

227. *See supra* Section III.A.

228. INSTITUTE OF MEDICINE, DISCUSSION FRAMEWORK FOR CLINICAL TRIAL DATA SHARING 13 (2014), [https://globalhealthtrials.tghn.org/site\\_media/media/medialibrary/2014/01/IOM\\_data\\_sharing\\_Report.pdf](https://globalhealthtrials.tghn.org/site_media/media/medialibrary/2014/01/IOM_data_sharing_Report.pdf) (stating that “shared clinical trial data might be analyzed in a manner that leads to biased effect estimates or invalid conclusions”).

229. Sally Kane, *12 Ways to Promote Your Blog: Blog Promotion Tips for Lawyers and Legal Professionals*, ABOUT.COM, <http://legalcareers.about.com/od/practicetips/tp/10-Ways-To-Promote-Your-Blog.htm>.

230. Daniel Vahab & Lisa Chau, *How to Promote Your Article Online*, SOCIAL MEDIA MONTHLY (Nov. 30, 2012), <http://thesocialmediamonthly.com/how-to-promote-your-article-online>.

231. Well-known examples are singing sensation Susan Boyle and conservative activist “Joe the Plumber.”

attention through word of mouth and social media, as happens when a YouTube video or blog post “goes viral.”<sup>232</sup>

While professional researchers most often seek publication in peer-reviewed journals that carefully scrutinize submissions, nothing will stop citizen scientists from posting their study results on blogs, personal web pages, and other electronic publications, making them instantaneously available to a worldwide audience.<sup>233</sup> Some commentators describe this phenomenon in terms of a shift from “intermediation” to “apomediation.”<sup>234</sup> Traditionally, peer reviewed journals served as necessary intermediaries between researchers and readers and thus gatekeepers for scientific knowledge. The Internet has now triggered disintermediation and increased use of apomediarities, agents or tools that guide readers to information without any middlemen required.<sup>235</sup> Many reports published on websites appear highly professional and credible to general readers, who are not always sophisticated about distinguishing between reliable and questionable sources of information.<sup>236</sup>

---

232. See Seth Mnookin, *One of a Kind: What Do You Do if Your Child Has a Condition That is New to Science?*, NEW YORKER (July 21, 2014), <http://www.newyorker.com/magazine/2014/07/21/one-of-a-kind-2> (describing how a father posted a blog entry about his disabled son’s extremely rare genetic abnormality in order to identify other patients with the condition, and the blog went viral, yielding contact with several other families).

233. R.J.W. Cline & K.M. Haynes, *Consumer Health Information Seeking on the Internet: The State of the Art*, 16 HEALTH EDUC. RES. 671, 679 (2001) (stating that “the Internet is characterized by uncontrolled and unmonitored publishing with little peer review”).

234. Dan O’Connor, *The Apomediated World: Regulating Research When Social Media Has Changed Research*, 41 J.L. MED. & ETHICS 470, 471 (2013); Gunther Eysenbach, *Medicine 2.0: Social Networking, Collaboration, Participation, Apomediation, and Openness*, 10 J. MED. INTERNET RES. e22 (2008) (coining the term “apomediation”).

235. Eysenbach, *supra* note 234, at 5.

236. See, e.g., Geraldine Peterson et al., *How Do Consumers Search For and Appraise Information on Medicines on the Internet? A Qualitative Study Using Focus Groups*, 5 J. MED. INTERNET RES. e33 (2003) (concluding “that there was a range of search and appraisal skills among [study] participants, with many reporting a limited awareness of how they found and evaluated Internet-based information on medicines”); Cline & Haynes, *supra* note 233, at 680 (cautioning that many consumers have weak information-evaluation skills); Miriam J. Metzger, *Making Sense of Credibility on the Web: Models for Evaluating Online Information and Recommendations for Future Research*, 58 J. AM. SOC’Y INFO. SCI. & TECH. 2078, 2079 (2007) (noting that “studies have found that users are seldom diligent in checking the accuracy of the information they obtain online”). *But see* S. Mo Jang, *Seeking Congruency or Incongruency Online? Examining Selective Exposure to Four Controversial Science Issues*, 36 SCI. COMM. 143, 159 (2014) (finding that “online users may not be as susceptible to confirmation bias [a tendency to favor information that confirms one’s views] as some scholars . . . have argued,” although “[t]hose who were more religious tended to avoid science news articles that challenged their existing views”).

Incorrect findings are unlikely to be a rarity. They will stem from a variety of failings and potentially lead to a number of different harms.

1. *Error Sources*

Erroneous findings could be caused by poor data quality in the original dataset or flawed study design.<sup>237</sup> Data quality deficiencies may result from clinicians' data entry errors in electronic health records, fragmented or incomplete electronic health records, data coding inaccuracies, or problems with software that processes or analyzes data.<sup>238</sup> Highly skilled analysts should be able to recognize data quality problems, adjust for them, and estimate error rates, but amateurs may not know how.<sup>239</sup>

Furthermore, scientific studies can be flawed due to a variety of biases. Selection bias arises when the group of subjects studied is not representative of the population as a whole, and thus, researchers cannot generalize study results.<sup>240</sup> For example, researchers using information from PatientsLikeMe or the Personal Genome Project should assume that individuals who choose to make their medical information public on such websites are a self-selected group (perhaps more educated and more interested in research) that is not typical of average patients. Confounding bias occurs when there are relevant variables that researchers neglect to consider that affect treatment choices and outcomes, and thus, the study's results are skewed.<sup>241</sup> For example, low income may be a confounder because it may cause individuals to select inferior, inexpensive treatments and may also separately lead to poor health because of stress or inadequate nutrition.<sup>242</sup> Measurement bias is a concern when measurements are inaccurate because equipment has failed, patients have reported facts incorrectly, or other problems have occurred in the process of collecting and measuring values.<sup>243</sup> Consequently, researchers face many hurdles and must conduct their studies very skillfully in order to derive valid results.

---

237. Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?*, 39 AM. J.L. & MED. 497, 515–27 (2013).

238. *Id.* at 515–21.

239. *Id.* at 530–32.

240. *Id.* at 521–23.

241. *Id.* at 523–25.

242. Sharona Hoffman & Andy Podgurski, *Big Bad Data: Law, Public Health, and Biomedical Databases*, 41 J.L. MED. & ETHICS (SUPPLEMENT ON 2012 PUB. HEALTH L. CONF.) 56, 58 (2013).

243. *Id.*

Researchers must be particularly sensitive to the difference between *association* and *causation*.<sup>244</sup> They may identify associations between certain behaviors, exposures, or treatments and particular outcomes but wrongly assume that there is a causal relationship between the two.<sup>245</sup> To illustrate, suppose that a citizen scientist concludes that people who eat acai berries live longer than those who do not eat this fruit. Does this mean that acai berry consumption prolongs life? Probably not. The explanation for this finding may well be that individuals who purchase this exotic fruit are generally well-off and have the means to make careful food choices, to exercise, to limit their stress, and to obtain top-notch medical care. Thus, it may be true that eating acai berries is *associated* with a longer life on average; but it does not follow that acai berries have some property that actually *causes* people to live longer.

Crowdfunding<sup>246</sup> may add another element of uncertainty to research quality. Crowdfunding does not depend on peer review of carefully written grant proposals by professional experts.<sup>247</sup> Rather, researchers aim to appeal to a large number of donors through videos and social media campaigns.<sup>248</sup> Some commentators have accused crowdfunding of turning “science into a popularity contest.”<sup>249</sup> It is certainly possible that the “crowd” will ignore the most meritorious proposals and opt to fund projects that are less deserving but more media-friendly and tantalizing.<sup>250</sup> Consequently, studies that are funded in this manner may not be of the highest quality.

## 2. *Potential Harms*

While many mistaken conclusions will be benign, some could be harmful. Patients reading incorrect information about their diseases may become unnecessarily anxious or, in the opposite case, overly sanguine about their symptoms and fail to seek needed medical care.

---

244. See, e.g., Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 PROC. ROYAL SOC'Y MED. 295, 295–300 (1965); Arvid Sjölander, *The Language of Potential Outcomes*, in CAUSALITY: STATISTICAL PERSPECTIVES AND APPLICATIONS 6, 9 (Carlo Berzuini et al. eds., 2012).

245. See Stephen Choi et al., *The Power of Proxy Advisors: Myth or Reality?* 59 EMORY L.J. 869, 879–85 (2010) (discussing the difference between correlation and causation).

246. See *supra* notes 113–115 and accompanying text.

247. Karen Kaplan, *Crowd-Funding: Cash on Demand*, 497 NATURE 147, 148 (2013).

248. *Id.*

249. Palca, *supra* note 115.

250. Kaplan, *supra* note 247, at 148.

Worse yet, individuals with personal agendas may undertake scientific studies with malevolent intent. They may use findings to inflame passion and prejudice against particular minority groups. Some may attempt to further political agendas by “proving” that their opponents’ policies have adverse effects on human health or the healthcare system. Others with selfish economic interests may aim to hurt competitors by claiming that their products cause particular ailments.<sup>251</sup>

Even peer-reviewed journals have published articles whose conclusions are false. A notorious example is a 1998 study published in the prestigious journal *Lancet*, that suggested a link between autism and the measles, mumps, rubella (MMR) vaccination.<sup>252</sup> While the study was later retracted,<sup>253</sup> the belief that vaccinations can lead to autism gained a considerable foothold and still needs to be explicitly repudiated on the CDC’s website.<sup>254</sup>

Researchers who are media-savvy or web-savvy and do not submit their findings to peer-reviewed journals for review by experts may be all the more likely to propagate incorrect and potentially harmful views. Manuscripts that are not submitted to journals will not be scrutinized by experts before their authors post them on the Internet, and no filtering mechanism exists to indicate to readers whether the material is valid or trustworthy.<sup>255</sup> The Internet provides publishing opportunities without any need for intermediaries and oversight. Therefore, potentially, millions of readers could view and believe even nonsensical conclusions, especially when authors assert that they based their research on data that the government furnished.

Many myths have in fact gained considerable traction despite the existence of abundant evidence to negate them. Two examples are climate change denial<sup>256</sup> and the outcry that the Patient Protection and Affordable

---

251. Michelle Mello et al., *Preparing for Responsible Sharing of Clinical Trial Data*, 369 *NEW ENG. J. MED.* 1651, 1653 (2013) (cautioning that public access to clinical trial data “could . . . lead unskilled analysts, market competitors, or others with strong private agendas to publicize poorly conducted analyses”).

252. Andrew J. Wakefield et al., *Ileal-Lymphoid-Nodular Hyperplasia, Non-Specific Colitis, and Pervasive Developmental Disorder in Children*, 351 *LANCET* 637, 641 (1998).

253. Simon H. Murch et al., *Retraction of an Interpretation*, 363 *LANCET* 750, 750 (2004).

254. *Measles, Mumps, and Rubella (MMR) Vaccine Safety Studies*, CDC, <http://www.cdc.gov/vaccinesafety/Vaccines/MMR/MMR.html> (last updated Aug. 28, 2015).

255. See *supra* notes 233–235 and accompanying text.

256. Aaron M. McCright & Riley E. Dunlap, *Cool Dudes: The Denial of Climate Change Among Conservative White Males in the United States*, 21 *GLOBAL ENVTL. CHANGE* 1163, 1163 (2011).

Care Act (aka Obamacare) would authorize “death panels” to decide which patients should live and which should die.<sup>257</sup> In both cases, the arguments gained popularity because high-profile public figures embraced them to further their own political agendas, which may occur in many other instances as well.

A particularly pernicious argument was made by Michael Levin in a 1997 book called *Why Race Matters*.<sup>258</sup> The author argued that African-Americans are typically less intelligent and more aggressive, assertive, and impulsive than Whites.<sup>259</sup> In addition, according to the author, African-Americans are more likely to commit crimes because they suffer from “an absence of conscience,” lack the ability to engage in self-monitoring, and have less free will and a different moral orientation from Whites.<sup>260</sup> In an era in which anyone in the world can access Internet material without leaving home or paying any money for a publication, these types of purportedly research-backed arguments can be more dangerous than ever before.

#### D. LITIGATION

Open health data may lead to a proliferation of litigation or threats of litigation in several circumstances. First, parties who feel they were injured by published invalid research outcomes may assert claims such as defamation or interference with economic advantage. Second, business entities may threaten to sue or file frivolous cases against citizen scientists who have acted in good faith and posted legitimate findings because the companies fear that the research outcomes will harm them. Thus, parties could use litigation to intimidate citizen scientists and pressure them to retract and remove purportedly offending materials. Third, data subjects who feel that they are victims of unauthorized disclosure of identifiable medical data may assert common law privacy breach claims. This Section analyzes several potential causes of action and the protection provided in some states by legislation that prohibits strategic lawsuits against public participation (SLAPPs).

---

257. Brian Beutler, *Republicans' "Death Panel" Smear Was Appallingly Effective*, NEW REPUBLIC (June 23, 2014), <http://www.newrepublic.com/article/118313/gop-obamacare-death-panel-smear-putting-peoples-lives-risk>.

258. MICHAEL LEVIN, *WHY RACE MATTERS: RACE DIFFERENCES AND WHAT THEY MEAN* (1997).

259. *Id.* at 213.

260. *Id.* at 213, 322.

### 1. Defamation

Defamation claims generally require proof of the following elements:

- (1) publication (to a third party)
- (2) of a defamatory statement
- (3) “of and concerning” the plaintiff
- (4) that is false,
- (5) published with requisite degree of fault (negligence or actual malice), and
- (6) damages the plaintiff’s reputation (which, in some instances, can be presumed).<sup>261</sup>

Establishing a successful defamation claim is no easy task, and plaintiffs must meet a high standard of proof.<sup>262</sup> Electronic speech is entitled to the same stringent First Amendment protections as print communication.<sup>263</sup>

Nevertheless, both individuals and entities may bring defamation claims.<sup>264</sup> For example, a manufacturer may file a defamation suit relating to the publication of intentionally false statements asserting that its product causes health problems. However, as a rule, defamatory statements against groups are not actionable.<sup>265</sup> Thus, if an author published or posted a piece asserting that Jews or African-Americans are biologically inferior in some way, Jewish or African-American plaintiffs could not bring a defamation claim, no matter how baseless and offensive the publication was.

An increasing number of defamation cases involve material posted on the Internet, which is the most likely venue for citizen science publications.<sup>266</sup> For example, businesses have filed defamation suits in response to negative reviews on the website Yelp.<sup>267</sup>

---

261. Matthew E. Kelley & Steven D. Zansberg, *A Little Birdie Told Me, “You’re A Crook”: Libel in the Twittersphere and Beyond*, 30 COMM. LAW. 34 (2014); RESTATEMENT (SECOND) OF TORTS § 558 (1977).

262. K.J. Greene, *Intellectual Property Expansion: the Good, the Bad, and the Right of Publicity*, 11 CHAP. L. REV. 521, 534 (2008) (stating that “defamation law sets very high standards of proof and injury to prevent conflict with First Amendment principles”).

263. *Reno v. ACLU*, 521 U.S. 844, 870 (1997) (asserting that “our cases provide no basis for qualifying the level of First Amendment scrutiny that should be applied to this medium [the Internet]”).

264. Wendy Gerwick Couture, *The Collision Between the First Amendment and Securities Fraud*, 65 ALA. L. REV. 903, 918–20 (2014) (discussing defamation suits brought by entities and individuals).

265. RESTATEMENT (SECOND) OF TORTS § 564A (1977); Ellyn Tracy Marcus, *Group Defamation and Individual Actions: A New Look at an Old Rule*, 71 CALIF. L. REV. 1532, 1533 (1983).

266. Amy Kristin Sanders & Natalie Christine Olsen, *Re-defining Defamation: Psychological Sense of Community in the Age of the Internet*, 17 COMM. L. & POL’Y 355, 365

A particularly memorable defamation case brought by industry involved a discussion on Oprah Winfrey's television show.<sup>268</sup> After scientists linked the consumption of beef from cattle infected by Mad Cow Disease with a new variant of the deadly Creutzfeldt-Jakob Disease, the *Oprah Winfrey Show*, like many other media outlets, covered the story in a segment entitled "Dangerous Foods."<sup>269</sup> At one point in the show Ms. Winfrey stated that she was "stopped cold from eating another burger."<sup>270</sup> Subsequently, several Texas cattlemen sued Ms. Winfrey and other defendants, asserting numerous causes of action, including defamation, and claiming that the beef market suffered significant losses because of the broadcast.<sup>271</sup> Fortunately for Oprah, the defendants prevailed on all claims.<sup>272</sup>

In some cases, plaintiffs may well have legitimate claims against individuals who maliciously publicize damaging information that they know to be false. In fact, the prospect of facing defamation claims may be an important deterrent to such misconduct. However, it is not difficult to imagine that in other instances, the chilling effect of litigation will thwart the dissemination of non-defamatory information. Industry may file lawsuits primarily to intimidate citizen scientists and force them to comply with demands for removal or retraction of material that they researched and posted in good faith. Citizen scientists who are far less powerful and prosperous than Oprah Winfrey may be unable to mount a full defense and simply capitulate.<sup>273</sup>

## 2. *Other Causes of Action*

Plaintiffs may file a myriad of other claims, only a few of which will be discussed as examples below. The cattle ranchers who sued Oprah Winfrey alleged not only defamation but also the closely related tort of business disparagement as well as negligence and negligence per se.<sup>274</sup> In addition, companies that feel their products have been inappropriately

---

(2012) (noting that "[w]ith the increasing number of speakers and messages has come a flurry of litigation as courts struggle to regulate the medium of the masses").

267. *Yelp, Inc. v. Hadeed Carpet Cleaning, Inc.* 752 S.E.2d 554 (Va. Ct. App. 2014); *Bently Reserve L.P. v. Papaliolios*, 160 Cal. Rptr. 3d 423 (2013).

268. *Texas Beef Group v. Winfrey*, 201 F.3d 680 (5th Cir. 2000).

269. *Id.* at 682–84.

270. *Id.* at 688.

271. *Id.* at 682.

272. *Id.* at 680.

273. *But see infra* Section IV.D.3 (discussing anti-SLAPP statutes).

274. *Texas Beef Group v. Winfrey*, 201 F.3d 680, 682 (5th Cir. 2000); *see id.* at 685 for a discussion of the elements of a business disparagement claim.

denigrated may bring a claim of interference with economic advantage. This theory of liability typically involves proof of the following elements: (1) plaintiff had an economic relationship with a third party that would have likely been economically beneficial for the plaintiff, (2) the defendant knew of the relationship, (3) the defendant engaged in intentional or negligent acts designed to disrupt the relationship, (4) the relationship was in fact disrupted, and (5) the defendant's conduct proximately caused plaintiff to suffer economic harm.<sup>275</sup> Individuals and entities subjected to published criticism or negative commentary often assert allegations of tortious interference with economic advantage alongside defamation claims.<sup>276</sup>

Patients whose data were used for research purposes may also initiate litigation. A patient who believes she did not consent to the posting of her identifiable medical records may assert a claim of public disclosure of private facts, a tort with the following elements: "(1) public disclosure (2) of a private fact (3) which would be offensive and objectionable to the reasonable person and (4) which is not of legitimate public concern."<sup>277</sup> There is no precedent for applying this theory of liability to re-identified data, but in the future, parties may attempt to invoke it in such circumstances. If re-identified medical information were posted on the Internet or otherwise publicized, the affected individuals may well find the conduct objectionable, and courts are likely to agree that the health records are not of public concern, thus ruling for plaintiffs.

### 3. *Anti-SLAPP Legislation*

Citizen scientists can take a degree of comfort in the existence of anti-SLAPP legislation in some states.<sup>278</sup> Strategic lawsuits against public participation (SLAPPs) have been defined as "civil complaints or counterclaims (against either an individual or an organization) in which the alleged injury was the result of petitioning or free speech activities protected by the First Amendment of the U.S. Constitution."<sup>279</sup> For

---

275. *Crown Imports, LLC v. Superior Court*, 223 Cal. App. 4th 1395, 1404 (2014) (discussing the tort under California law).

276. *Responding to Strategic Lawsuits Against Public Participation (SLAPPs)*, DIGITAL MEDIA LAW PROJECT, <http://www.dmlp.org/legal-guide/responding-strategic-lawsuits-against-public-participation-slapps> (last visited Nov. 23, 2015).

277. *See Diaz v. Oakland Tribune, Inc.*, 139 Cal. Rptr. 762, 768 (Cal. Ct. App. 1983) (listing the elements of the public disclosure tort under California law).

278. DIGITAL MEDIA LAW PROJECT, *supra* note 276.

279. Robert D. Richards, *A SLAPP in the Facebook: Assessing the Impact of Strategic Lawsuits against Public Participation on Social Networks, Blogs, and Consumer Gripe Sites*, 21 DEPAUL J. ART, TECH. & INTELL. PROP. L. 221, 222 (2011).

example, SLAPPs have been filed by businesses as a form of retaliation against consumers who posted negative comments about them on social networking sites.<sup>280</sup> There is thus reason to worry that some companies will file SLAPPs against citizen scientists who claim that their products are inferior to others or cause health-related harms.

Anti-SLAPP statutes have been enacted in twenty-eight states, the District of Columbia, and Guam.<sup>281</sup> These laws enable defendants subject to certain frivolous allegations to have SLAPPs dismissed quickly and to recover costs and attorneys' fees.<sup>282</sup> The statutes can vary significantly.<sup>283</sup> Pennsylvania's law is very narrow, granting immunity to defendants who make "an oral or written communication to a government agency relating to enforcement or implementation of an environmental law or regulation . . . ."<sup>284</sup> By contrast, in California the law is much broader and covers "written or oral statement[s] or writing made in a place open to the public or a public forum in connection with an issue of public interest."<sup>285</sup> The Pennsylvania law allows defendants to request hearings at which the court will determine whether they are entitled to immunity.<sup>286</sup> The California law establishes a somewhat different procedure, allowing a covered defendant to file a special motion to strike, after which the court will require the plaintiff to produce evidence that it is likely to prevail on its claim. In the absence of such evidence, the claim will be dismissed and defendant will recover attorney's fees and costs.<sup>287</sup> Protection is inconsistent across jurisdictions but may be very helpful to some victims of frivolous litigation initiated for purposes of harassment and intimidation.

## V. RECOMMENDATIONS

The growing trend of opening patient-related data held by the government and private entities to the public raises hopes for considerable benefits. At the same time, it provokes significant concerns. How should

---

280. *Id.* at 222–23; Rex Hall, Jr., *Firm Sues WMU Student Over Facebook Page; Towing Company Seeks \$750,000 in Damages for Online Criticism*, GRAND RAPIDS PRESS, Apr. 14, 2010, at A6 (discussing litigation that followed the student's posting of an entry on his Facebook page that criticized T & J Towing for wrongly towing his car from a legal parking space and damaging it).

281. DIGITAL MEDIA LAW PROJECT, *supra* note 276.

282. *Id.*

283. Richards, *supra* note 279, at 232.

284. 27 PA. CONS. STAT. ANN. § 8302(a) (2001).

285. CAL. CODE CIV. P. §§ 425.16(e), 425.17 (West 2011).

286. 27 PA. CONS. STAT. ANN. § 8303 (2001).

287. CAL. CODE CIV. P. § 425.16(b)–(c) (2011).

legislators and regulators respond to this emerging phenomenon? The law must balance the interests of a variety of stakeholders: patients, professional researchers, citizen scientists, government, industry, and the public at large. An excessively heavy-handed approach to regulation might discourage citizen scientists from pursuing projects and making important contributions and may deter data custodians from releasing records. However, a regulatory approach that is too timid may result in privacy breaches, discrimination, and other societal harms. This Part formulates recommendations for regulatory and policy modifications to address open data concerns.

#### A. PRIVACY AND DATA STEWARDSHIP

The risk that anonymized health information will be re-identified and used inappropriately can never be fully eliminated,<sup>288</sup> but it can be minimized. Several legal and policy interventions could enhance privacy protections. First, the HIPAA Privacy Rule should be amended to expand the definition of “covered entity” and to add a provision that prohibits re-identification. Second, any party releasing patient-related data to the public should establish a data release review board that will scrutinize all disclosed data sets to ensure that they are de-identified as effectively as possible. The review board should also oversee other privacy protections, including privacy training for data recipients, data use agreements, user registries, and consent procedures for data subjects opting to share identifiable information.

##### 1. *HIPAA Privacy Rule Modifications*

Two HIPAA Privacy Rule changes should be made to enhance data subject privacy. The HIPAA statute and regulations should be amended to expand their reach and efficacy through a broader definition of “covered entity” and an explicit prohibition of any attempt to re-identify data.

##### a) Expanding the Definition of “Covered Entity” and Creating National Data Release and De-identification Standards

The HIPAA Privacy Rule currently governs only healthcare providers, health plans, healthcare clearinghouses, and their business associates.<sup>289</sup> It therefore does not apply to numerous parties that store and disclose health information, including government entities and database operators. Expansion of the definition of “covered entity” in the HIPAA Privacy

---

288. *See infra* Section IV.A.3.

289. 45 C.F.R. §§ 160.102–.103 (2013); 42 U.S.C. § 17934 (2010).

Rule and its enabling legislation<sup>290</sup> could improve privacy protection for data subjects. Regulators could turn to a Texas privacy statute as a model for more comprehensive coverage. The law defines “covered entity” in relevant part as any party who,

for commercial, financial, or professional gain, monetary fees, or dues, or on a cooperative, nonprofit, or pro bono basis, engages, in whole or in part, and with real or constructive knowledge, in the practice of assembling, collecting, analyzing, using, evaluating, storing, or transmitting protected health information. The term includes a business associate, health care payer, governmental unit, information or computer management entity, school, health researcher, health care facility, clinic, health care provider, or person who maintains an Internet site.<sup>291</sup>

The HIPAA Privacy Rule’s scope of coverage should be similarly broadened, with one modification. The regulations should explicitly reach employers, financial institutions, and amateur researchers, along with the parties listed in the definition above.

The proposed regulatory expansion should not inhibit the release of data to the public. Rather, it would provide all data holders with clear instructions regarding privacy safeguards and create uniform, national standards for data disclosure and de-identification.<sup>292</sup> Those releasing identifiable information, such as PatientsLikeMe or the Personal Genome Project would need to obtain meaningful patient consent,<sup>293</sup> as discussed in greater detail below.<sup>294</sup> Those who wish to be exempt from HIPAA coverage would need to de-identify disclosed data in accordance with the Privacy Rule’s de-identification provision.<sup>295</sup>

In some cases, data holders will want to release information that is largely anonymized but contains a few identifiers that are particularly useful for research purposes. In these instances, database operators would follow the Privacy Rule’s “limited data set” provision.<sup>296</sup> In limited data

---

290. 45 C.F.R. §160.103 (2013); 42 U.S.C. §1320d-1(a) (2010).

291. TEX. HEALTH & SAFETY CODE ANN. 181.001(b)(2)(A) (West 2012).

292. Note that the definition of “health information” would also need to be revised because it is currently limited to information that is “created or received by a health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse.” 45 C.F.R. § 160.103 (2013). It thus fails to include data handled by website operators and others.

293. 45 C.F.R. §164.508 (2013).

294. See *infra* notes 318–321 and accompanying text.

295. 45 C.F.R. §164.514(b)(2013); See *supra* Section IV.A.2 for detailed discussion of de-identification.

296. 45 C.F.R. §164.514(e)(1)–(4) (2013).

sets, custodians redact most of the safe harbor provision's eighteen identifiers but retain dates and geographic locales, including city or town, state, and postal codes.<sup>297</sup> Database operators may release limited data sets without patient authorization so long as data recipients sign data use agreements containing specified restrictions and privacy protections.<sup>298</sup> These agreements are required because the added identifiers, while valuable to analysts, make re-identification considerably easier for skilled attackers.<sup>299</sup>

The proposed change would modify only the definition of "covered entity." It would not impact the exceptions to the HIPAA Privacy Rule that the regulations establish elsewhere.<sup>300</sup> Thus, the proposal would not create hurdles for health care treatment, payment, administration, or the activities of law enforcement and public health officials.<sup>301</sup>

b) Prohibiting Re-identification

The HIPAA Privacy Rule should also be amended to include a general prohibition of any attempt to re-identify information that would apply to any user of de-identified data.<sup>302</sup> This restriction is already an element of data use agreements, which require the recipients of limited data sets to promise that they will not "identify the information or contact the individuals."<sup>303</sup> The proposed change would extend this regulatory proscription to anyone using de-identified information, including employers, financial institutions, and all other parties. The provision could specify exceptions, such as permitting re-identification necessary to respond to medical or public health emergencies. Violators should be subject to HIPAA's enforcement provisions, which incorporate civil and criminal penalties.<sup>304</sup>

---

297. 45 C.F.R. §164.514(e)(2) (2013).

298. 45 C.F.R. §164.514(e)(4) (2013).

299. Kathleen Benitez & Bradley Malin, *Evaluating Re-identification Risks with Respect to the HIPAA Privacy Rule*, 17 J. AM. MED. INFORMATICS ASS'N 169, 169 (2010) (estimating that the risk of re-identification is between 10% and 60%, depending on the state).

300. 45 C.F.R. §§ 164.502, .506, .512 (2013).

301. *Id.*

302. If the HIPAA Privacy Rule's scope of coverage is expanded as suggested above, the prohibition would apply to all covered entities and individuals. If not, the HIPAA statute itself should be amended to include a re-identification prohibition that applies broadly to all de-identified health data users.

303. 45 C.F.R. §164.514(e)(4)(ii)(C)(5) (2013).

304. 45 C.F.R. §§ 160.300-.552 (2013).

## 2. *Data Release Review Boards*

In the absence of HIPAA Privacy Rule amendments, data custodians not currently covered by the Rule should implement their own privacy safeguards. Database operators who release patient-related information to the public should institute a thoughtful and thorough process for reviewing the information at issue and establishing strong privacy safeguards.

The CDC's Policy on Releasing and Sharing Data recommends the establishment of data-release review boards, and data custodians would be wise to implement this suggestion.<sup>305</sup> The boards, composed of data mining and privacy experts, would review any data that are to be released to ascertain that they are as effectively de-identified as possible. For example, the board would assess whether the disclosed sample size is so small that data subjects are likely to be identified no matter what variables are stripped away, as may be the case when data is collected about very rare diseases.<sup>306</sup> The board would also determine what statistical methods should be used to achieve de-identification of various data sets, including suppression, perturbation, and generalization.<sup>307</sup> In addition, the board could analyze data quality to ensure that the released information is sufficiently reliable that it will be of value to users.<sup>308</sup> Finally, the data-release review board should oversee all other privacy safeguards that data holders implement.

## 3. *Data Use Agreements, Privacy Training, Registries, and Consent Procedures*

Data custodians who release medical information to the public should implement several privacy protection measures beyond board review, and the extent of these procedures should depend on the type of data at issue. Users who access any database of medical information, including aggregate, summary-level data, should be alerted that the information is sensitive and raises privacy concerns. For example, the CDC Wonder website asks viewers who are seeking mortality information from its

---

305. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 9.

306. *See supra* note 160 and accompanying text.

307. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 9; *supra* note 157 and accompanying text (discussing the various statistical methods).

308. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 9; *see supra* Section IV.C.1 (discussing data quality shortcomings); Hoffman & Podgurski, *supra* note 237, at 530–32 (discussing data quality assessment).

database to agree to a short list of data use restrictions by clicking an “I agree” icon.<sup>309</sup>

For patient-level data that is not aggregated, more elaborate procedures are needed. The Healthcare Cost and Utilization Project’s National (Nationwide) Inpatient Sample (NIS) furnishes a useful model. The NIS contains information concerning millions of hospital stays.<sup>310</sup> It provides detailed information about patients and hospitals but is careful to remove identifiers and most likely meets the HIPAA safe harbor standard.<sup>311</sup> Nevertheless, the NIS requires purchasers of the data to take a 15-minute training course that addresses privacy concerns.<sup>312</sup> It also requires purchasers to sign a detailed data use agreement that specifies a variety of use restrictions designed to protect individual and institutional data subjects from privacy violations and other abuses, such as attempts to

---

309. *About Underlying Cause of Death, 1999–2013*, CDC WONDER, <http://wonder.cdc.gov/ucd-icd10.html>. Users agree that they will:

- Use these data for health statistical reporting and analysis only.
- For sub-national geography, do not present or publish death counts of 9 or fewer or death rates based on counts of nine or fewer (in figures, graphs, maps, tables, etc.).
- Make no attempt to learn the identity of any person or establishment included in these data.
- Make no disclosure or other use of the identity of any person or establishment discovered inadvertently and advise the NCHS Confidentiality Officer of any such discovery.

*Id.*

310. *Overview of the National (Nationwide) Inpatient Sample (NIS)*, HEALTHCARE COST & UTILIZATION PROJECT, <http://www.hcup-us.ahrq.gov/nisoverview.jsp> (last updated Nov. 17, 2015).

311. *Id.* The data elements provided in the overview are:

- Primary and secondary diagnoses and procedures;
- Patient demographic characteristics (e.g., sex, age, race, median household income for ZIP Code);
- Hospital characteristics (e.g., ownership);
- Expected payment source;
- Total charges;
- Discharge status;
- Length of stay;
- Severity and comorbidity measures.

*Id.*

312. *Welcome to the HCUP Data Use Agreement (DUA) Training!*, HEALTHCARE COST & UTILIZATION PROJECT, [http://www.hcup-us.ahrq.gov/tech\\_assist/dua.jsp](http://www.hcup-us.ahrq.gov/tech_assist/dua.jsp) (last updated Sept. 23, 2015).

gain commercial or competitive advantage through analysis of released NIS data.<sup>313</sup>

If data users violate the agreement, the NIS would presumably challenge them in court.<sup>314</sup> A useful supplement to the NIS's requirements would be an online test in which examinees would have to demonstrate that they read and understood the training materials and data use agreement.

Admittedly, training courses and data use agreements will not prevent all privacy violations, and data custodians are not likely to dedicate significant resources to their enforcement. However, these measures will alert the public to the importance of privacy and responsible data handling and may avert innocent breaches by citizen scientists who wish to do no harm.

Equally important, the data use agreement requirement will create a record of those accessing data, and data custodians should maintain functional registries of users. Data custodians can require signatories to provide their name, affiliation, and contact information.<sup>315</sup> If the dataset at issue consists of lower-risk, aggregated or summary data and users do no more than click on an "I agree" icon, only their network addresses will be recorded. Nevertheless, if the individuals used their own computers, authorities could link the network addresses to their identities if need be.<sup>316</sup> Data custodians could then preclude those who violate data use agreements by re-identifying data or engaging in other misconduct from downloading information in the future, and the government could subject such violators to other penalties.<sup>317</sup>

In some cases, privacy requirements should apply not only to data users, but also to data subjects. Specifically, individuals choosing to allow

---

313. HEALTHCARE COST & UTILIZATION PROJECT, DATA USE AGREEMENT FOR THE NATIONWIDE DATABASES (2014), [http://www.hcup-us.ahrq.gov/team/HCUP\\_Nationwide\\_DUA\\_051614.pdf](http://www.hcup-us.ahrq.gov/team/HCUP_Nationwide_DUA_051614.pdf).

314. *See id.* (explaining that violation of the data use agreement can lead to fines or imprisonment under federal and state law); *About Underlying Cause of Death*, *supra* note 309 (describing sanctions for violations).

315. *Data Use Agreement*, *supra* note 313.

316. J.D. Sartain, *Can Your IP Address Give Away Your Identity to Hackers, Stalkers and Cybercrooks?*, NETWORKWORLD (Jul. 16, 2013), <http://www.networkworld.com/article/2168144/malware-cybercrime/can-your-ip-address-give-away-your-identity-to-hackers--stalkers-and-cybercrooks-.html>. Devious persons may, however, use a spoofed Internet address.

317. *About Underlying Cause of Death*, *supra* note 309 (describing sanctions for violations and stating that "[r]esearchers who violate the terms of the data use restrictions will lose access to WONDER and their sponsors and institutions will be notified").

public access to identifiable or easily identifiable data, such as datasets that include birth date, sex, and zip code,<sup>318</sup> should undergo a comprehensive informed consent process.<sup>319</sup> Such data subjects should understand that their personal health information will be viewable not only by researchers with good intentions, but also by employers, marketers, financial institutions, and others who may not have their best interest in mind.<sup>320</sup> To this end, the Harvard Personal Genome Project requires participants to read and sign a lengthy consent document. They also must pass an examination demonstrating their understanding of the material contained in the consent form.<sup>321</sup> Testing data subjects' comprehension of the privacy risks they accept would be an important component of any informed consent process pertaining to sharing individually identifiable data.

#### B. ANTI-DISCRIMINATION PROTECTIONS

Ironically, while open data policies promote transparency on the government's part,<sup>322</sup> they may provide new opportunities for employers and others to discriminate in non-transparent ways.<sup>323</sup> Based on data about various health risks, entities might discriminate against discrete population sub-groups such as African-American women older than fifty.<sup>324</sup> These multi-factor discrimination cases are much more difficult to detect and prosecute than cases involving traditional protected classes.<sup>325</sup> In addition, entities may retain experts to mine data and develop new applicant

---

318. See El Emam, *supra* note 163, and accompanying text (explaining that it is relatively easy to re-identify such data).

319. Arguably, anyone whose data is released to the public in any form, including as fully de-identified information, should be asked for consent. A full exploration of this issue is beyond the scope of this Article. However, it is unrealistic to expect that government authorities who receive data relating to millions of patients from a variety of sources will have the resources to track down, contact, and obtain consent from all data subjects. Moreover, allowing individuals to opt out of data sharing could lead to selection bias, whereby the people who choose to be included in databases are not representative of the population as a whole. If that is the case, research results based on study of database participants could not be generalized to others, and therefore, would be of very limited scientific use. Therefore, this Article recommends extensive consent procedures only for data subjects who opt to disclose identifiable or easily re-identifiable information. See Hoffman & Podgurski, *supra* note 15, at 114–123 (discussing the problems with consent).

320. See *supra* Section IV.B (discussing discrimination concerns).

321. *Participation Documents*, PERSONAL GENOME PROJECT, HARV. MED. SCH., <http://www.personalgenomes.org/harvard/sign-up#documents> (last visited Nov. 23, 2015).

322. See *supra* notes 123–124 and accompanying text.

323. See *supra* Section IV.B.

324. See *supra* notes 203–204 and accompanying text.

325. See *supra* notes 207–208 and accompanying text.

screening tools that focus on proxies for disability or predictors of bad health that employers can consider without violating any explicit legal prohibition.<sup>326</sup> As open data and data mining proliferate, novel forms of health-based discrimination may become increasingly common and require several changes to anti-discrimination law and practice.

1. *Detecting, Deterring, and Prosecuting Multi-Factor Discrimination*

As difficult as multi-factor discrimination may be to detect, enforcement agencies and plaintiffs' attorneys will need to recognize the real possibility that it is occurring.<sup>327</sup> An uptick in litigation and enforcement actions relating to multi-factor cases may encourage victims to bring this type of discrimination to light and discourage employers and businesses from engaging in it.

In multi-factor cases, plaintiffs claiming employment discrimination who believe that one of the improperly considered attributes was their age may face particular hurdles because of the Supreme Court's decision in *Gross v. FBL Financial Services, Inc.* This decision barred mixed-motive claims and required "but for" proof of age discrimination.<sup>328</sup> However, in *Gross* the employer allegedly considered a mixture of proper (performance-related) and improper (the plaintiff's age of fifty-four) factors rather than a combination of prohibited categories (for example, age, race, sex).<sup>329</sup> In a future case, the Supreme Court may revisit the question of whether plaintiffs can sue employers for discriminating based on age and one or more other protected classifications and hold that such claims are allowable. In the alternative, Congress could amend the Age Discrimination in Employment Act to add a provision that explicitly permits multi-factor claims.<sup>330</sup>

---

326. See *supra* notes 211–218 and accompanying text.

327. See *supra* note 208 and accompanying text; Cathy Scarborough, *Conceptualizing Black Women's Employment Experiences*, 98 YALE L.J. 1457, 1476–78 (1989) (discussing Title VII multi-factor claims).

328. Day, *supra* note 204, at 466–67; *Gross v. FBL Fin. Servs., Inc.*, 557 U.S. 167, 177–78 (2009).

329. FBL's defense was that "Gross' reassignment was part of a corporate restructuring and that Gross' new position was better suited to his skills" and no protected classification other than age was at issue. *Gross*, 557 U.S. at 167.

330. See Day, *supra* note 204, at 466–67 (proposing legislative action to approve age-plus-sex claims).

2. *Requiring Disclosure of Data Mining for Disability Proxies and Predictors*

Instances in which employers, financial institutions, or others engage in data mining and exclude individuals based on perceived or anticipated health conditions will also be difficult to detect. Consequently, anti-discrimination laws should include a requirement that businesses disclose their data mining practices to workers, consumers, and other parties that are affected by such practices.

Several other commentators have called for transparency with respect to data mining and predictive modeling activities. Professors Danielle Citron and Frank Pasquale argue that “we need to switch the default in situations like this away from an assumption of secrecy, and toward the expectation that people deserve to know how they are rated and ranked.”<sup>331</sup> Similarly, commentators Kate Crawford and Jason Schultz would require parties to provide notice, “disclosing not only the type of predictions they attempt, but also the general sources of data that they draw upon as inputs, including a means whereby those whose personal data is included can learn of that fact.”<sup>332</sup>

A disclosure requirement would be a valuable addition to anti-discrimination protections. It would constitute a compromise between prohibiting data mining practices altogether and ignoring them. A tweak of the ADA’s medical inquiry and exam provision<sup>333</sup> could add a requirement that employers disclose in writing to applicants and employees any medical data mining activities that they intend to use for purposes of making employment decisions. This information would then be available to plaintiffs’ attorneys and government enforcement agencies such as the Equal Employment Opportunity Commission (EEOC),<sup>334</sup> which could investigate whether these activities resulted in unlawful discrimination. Likewise, the ADA’s public accommodation title could feature the same provision to cover financial institutions and other businesses.<sup>335</sup> Employment or loan application forms could include disclosure statements so long as the statements were in sufficiently large and readable print or on separate sheets given to applicants.

---

331. Citron & Pasquale, *supra* note 185, at 21.

332. Crawford & Schultz, *supra* note 185, at 125.

333. 42 U.S.C. § 12112(d) (2010).

334. The Equal Employment Opportunity Commission is the federal agency tasked with enforcing the federal anti-discrimination laws. See *Overview*, EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, <http://www.eeoc.gov/eeoc> (last visited Nov. 23, 2015).

335. See 42 U.S.C. § 12182 (2010).

Some may object that such a requirement will open the floodgates of litigation, especially in employment discrimination cases, because any applicant who receives notice of an employer's data mining activities and who is not hired or promoted could claim discrimination. Employment discrimination claimants, however, must exhaust their administrative remedies prior to filing suit.<sup>336</sup> While the EEOC and state administrative agencies would likely be able to hire experts to investigate and interpret employers' data mining activities in selected instances, they pursue litigation in only a handful of cases each year because of limited resources.<sup>337</sup> The vast majority of claimants, whose cases the government will not pursue, will need to find an attorney interested in investing the time and money in delving into the technicalities of data mining activities, which may be no easy task.<sup>338</sup> Furthermore, plaintiffs would have legitimate claims only if they were subjected to discrimination based on legally protected characteristics such as race or disability. Still, the existence of a disclosure requirement may deter at least some employers from engaging in unlawful discrimination and depriving qualified employees of job opportunities.

### 3. *Addressing Data Mining in the ADA's Definition of Disability*

The ADA defines "disability" very broadly<sup>339</sup> and prohibits employers, financial institutions, and others from discriminating against individuals based on a belief that they currently have physical or mental impairments. The ADA's "regarded as" provision explicitly states that an individual is protected by the statute if "he or she has been subjected to an action prohibited under this chapter because of an actual or perceived physical or mental impairment whether or not the impairment limits or is perceived to limit a major life activity."<sup>340</sup>

---

336. 42 U.S.C. § 2000e-5(e)-(f), 42 U.S.C. § 12117 (addressing EEOC enforcement responsibilities). Title III of the ADA, which covers public accommodations such as financial institutions does not include a similar requirement that plaintiffs exhaust administrative remedies. *See Hill v. Park*, No. 03-4677, 2004 WL 180044 (E.D. Pa. Jan. 27, 2004).

337. *See EEOC Litigation Statistics*, EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, <http://www.eeoc.gov/eeoc/statistics/enforcement/litigation.cfm> (last visited Nov. 23, 2015) (indicating that in fiscal year 2013, the EEOC filed only 148 lawsuits nationwide).

338. *See* Theodore J. St. Antoine, *Mandatory Arbitration: Why It's Better than It Looks*, 41 U. MICH. J.L. REFORM 783, 790 (2008) (estimating that only 5% of individuals with employment discrimination claims who turn to private attorneys for help are actually able to retain counsel).

339. *See* 42 U.S.C. § 12102 (2010).

340. 42 U.S.C. § 12102(3)(A) (2010).

However, the ADA does not ban discrimination against individuals who are neither currently impaired nor perceived as impaired but are deemed to be at risk of being unhealthy in the future because of their eating habits, exposure to toxins, or a myriad of other concerns.<sup>341</sup> Thus, for example, so long as employers do not consider genetic factors,<sup>342</sup> they can exclude such workers without being challenged.

If open data enables discrimination against high health-risk individuals and such discrimination becomes increasingly common, legislators would be wise to respond to it. An easy fix would be to add language to the “regarded as” provision of the ADA indicating that individuals are also regarded as disabled if they have been subjected to an adverse action because they are perceived as likely to develop physical or mental impairments in the future.

### C. CITIZEN SCIENTIST CHAPERONING

Several mechanisms should be developed to assist citizen scientists in conducting, validating, and publishing their research. “Chaperoning” citizen scientists by means of research support and filtering tools could reduce the potential for widespread dissemination of erroneous and harmful research conclusions.<sup>343</sup>

First, government agencies, academic institutions, and other research experts should develop educational resources and best practices guidelines to assist citizen scientists in conducting research.<sup>344</sup> These documents or videos could be posted on database websites, and users could be required or encouraged to review them, along with privacy training materials, before signing data use agreements.<sup>345</sup> Data custodians could also test users

---

341. *See id.*

342. *See* Hoffman, *supra* note 218 and accompanying text (discussing the Genetic Information Nondiscrimination Act).

343. *See supra* Section IV.C.

344. CDC/ATSDR POLICY ON RELEASING AND SHARING DATA, *supra* note 111, at 7 (urging CDC staff to develop “[i]nstructions for non-CDC users on the appropriate use of the data”); JOHN P. HOLDREN, OFFICE SCI. & TECH. POL’Y, EXEC. OFFICE OF THE PRESIDENT, MEMORANDUM, INCREASING ACCESS TO THE RESULTS OF FEDERALLY FUNDED SCIENTIFIC RESEARCH 6 (2013) [https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp\\_public\\_access\\_memo\\_2013.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf) (urging federal agencies, in coordination with the private sector, to “support training, education, and workforce development related to scientific data management, analysis, storage, preservation, and stewardship”).

345. *See supra* Section V.A.3.

on these materials in order to ensure that they have read and understood them prior to signing the agreement.<sup>346</sup>

Second, citizen scientists should have opportunities to have their work vetted, validated, and published in platforms that are recognized as reliable. Without such mechanisms, readers will be unable to discern whether citizen scientists' findings are trustworthy.

One option is to follow the Wikipedia paradigm. Wikipedia allows the public to post and edit articles, but the site provides some degree of oversight and quality control.<sup>347</sup> Authors can request reviews of their entries from peers, and Wikipedia administrators have authority to delete and undelete pages, protect pages from editing, and take other actions.<sup>348</sup> In extreme cases, administrators, of whom there are over 1,400, can temporarily or permanently bar authors from contributing to Wikipedia because of intentional and persistent misconduct.<sup>349</sup> In addition, Wikipedia has an extensive dispute resolution system for disagreements about the contents of Wikipedia pages.<sup>350</sup> Wikipedia encourages readers who find passages that are biased or erroneous to improve them and discuss the problem with the original author. Parties may also ask for a third opinion or for a moderated discussion through the Dispute Resolution Noticeboard, or they can initiate open requests for comments from the community at large or requests for mediation with help from the Mediation Committee.<sup>351</sup>

A similar venue could be established for the publication of citizen scientists' reports and findings that are not submitted to traditional journals. Opportunities for editing by other professional and amateur scientists, dispute resolution mechanisms, and other forms of oversight would significantly enhance the reliability of posted materials. The venue's policy should also require authors to disclose any computer programs that

---

346. See *Participation Documents*, *supra* note 321 and accompanying text.

347. *Policies and Guidelines*, WIKIPEDIA (Feb. 5, 2016, 11:49 AM), [https://en.wikipedia.org/w/index.php?title=Wikipedia:Policies\\_and\\_guidelines&oldid=703418205](https://en.wikipedia.org/w/index.php?title=Wikipedia:Policies_and_guidelines&oldid=703418205).

348. *Editor Review*, WIKIPEDIA (Dec. 9, 2015, 2:24 PM) [https://en.wikipedia.org/w/index.php?title=Wikipedia:Editor\\_review&oldid=694475551](https://en.wikipedia.org/w/index.php?title=Wikipedia:Editor_review&oldid=694475551); *Administrators*, WIKIPEDIA (Jan. 28, 2016, 2:56 PM), <https://en.wikipedia.org/w/index.php?title=Wikipedia:Administrators&oldid=70211392>.

349. *Id.*; *Policies and Guidelines*, *supra* note 347.

350. *Dispute Resolution*, WIKIPEDIA (Oct. 26, 2015, 4:16 PM), [https://en.wikipedia.org/w/index.php?title=Wikipedia:Dispute\\_resolution&oldid=687601794](https://en.wikipedia.org/w/index.php?title=Wikipedia:Dispute_resolution&oldid=687601794).

351. *Id.*

they used to analyze the data so that others can replicate and verify their research.<sup>352</sup>

Opportunities for peer review of citizen science research outcomes would provide significant benefits. The contemporary scientific community is open to innovation and several hybrid peer review models are emerging. For example, F1000Research is a pioneering open access journal for life scientists.<sup>353</sup> F1000Research reviews submitted articles internally and, if it initially deems them meritorious, it publishes them within a week of submission, together with their underlying datasets, making all materials publicly available. The service only then sends articles for peer review. Another novelty is that F1000Research discloses its reviewers' identities and enables authors to communicate with the reviewers to address their concerns. Authors may publish revised manuscripts,<sup>354</sup> and articles that peer reviewers approve are indexed in external databases such as PubMed.<sup>355</sup>

Peerage of Science offers a second non-traditional approach.<sup>356</sup> Authors submit manuscripts to the service rather than directly to journals. Authors set their own deadlines for reviews, and any qualified reviewer with a prior peer-reviewed publication can submit a review. A second stage of the process reviews the initial reviewers' assessments.<sup>357</sup> Authors can accept offers from participating journals or export reviews outside of Peerage of Science to journals of their choice.<sup>358</sup>

F1000Research and Peerage of Science demonstrate the contemporary spirit of innovation in the academic community. They are not suggested as venues for amateur citizen scientists, because they are designed for professional scientists producing conventional scholarship. The future, however, may herald different models to chaperone citizen scientists. Whether these follow the Wikipedia paradigm or another path, they would assist not only researchers in improving and publicizing their work,

---

352. Ari B. Friedman, Letter to the Editor, 370 *NEW. ENG. J. MED.* 484, 484 (2014) (reviewing Michelle M. Mello et al., *Preparing for Responsible Sharing of Clinical Trial Data*, 369 *NEW. ENG. J. MED.* 1651 (2013)).

353. *How It Works*, F1000RESEARCH, <http://f1000research.com/about> (last visited Nov. 23, 2015).

354. *Id.*

355. *FAQs*, F1000RESEARCH, <http://f1000research.com/faqs> (last visited Nov. 23, 2015).

356. *How It Works*, PEERAGE OF SCIENCE, <http://www.peerageofscience.org/how-it-works> (last visited Nov. 23, 2015).

357. *Process Flow*, PEERAGE OF SCIENCE, <http://www.peerageofscience.org/how-it-works/process-flow> (last visited Nov. 23, 2015).

358. *Id.*

but also the reading public in filtering out research findings that have no reliable basis.<sup>359</sup>

#### D. TORT CLAIM LITIGATION STRATEGIES

Parties who are hurt by citizen scientists' wrongdoing will have a variety of avenues to seek redress. Plaintiffs may allege defamation, interference with economic advantage, public disclosure of private facts, and other claims.<sup>360</sup> Database operators who require data recipients to sign data use agreements may also sue for breach of contract if (1) recipients attempt to re-identify information, use data for commercial or competitive purposes, or violate other agreement provisions, and (2) the breaches damage the database's reputation or economic interests.<sup>361</sup>

Of greater concern are instances in which parties may file suit against citizen scientists who act in good faith but publicize information critical of the plaintiffs' products or conduct. Businesses may hope to intimidate and deter citizen scientists and to force them to disavow and remove any offending material.<sup>362</sup> Citizen scientists who publish their data outside of traditional academic journals will not have a defense based on scrutiny and approval by highly qualified peer reviewers. Such citizen scientists will have no academic institution committed to their vigorous defense.

In some states, defendants will be able to utilize anti-SLAPP legislation and have cases quickly dismissed.<sup>363</sup> If amateur researchers make valuable contributions to science but are routinely harassed through frivolous litigation, additional states may respond with anti-SLAPP statutes that cover such cases.

In the meantime, citizen scientist advocacy organizations can develop educational materials that address strategies to minimize the risk of liability. To this end, the Harvard-affiliated Digital Media Law Project offers "Practical Tips for Avoiding Liability Associated with Harms to

---

359. Admittedly, even experienced scientists often cannot reach consensus about the validity of research findings and disagree about the accuracy of study outcomes. *See supra* notes 98–102 and accompanying text. However, a filtering mechanism could at least screen out material that no educated reviewer would consider reliable.

360. *See supra* Sections IV.D.1 and IV.D.2.

361. *See supra* notes 309–313 and accompanying text.

362. *See supra* Section IV.D.3.

363. *Id.*

Reputation.”<sup>364</sup> The long list of detailed suggestions includes, among others:

- Strive to be as accurate as possible;
- Use reliable sources;
- Seek comment from the subjects of your statements, when appropriate;
- Document your research;
- Keep an eye out for “Red Flag” statements [e.g., explicitly accusing someone of criminal or immoral conduct];
- Be cautious when publishing negative information about businesses;
- Where possible, get consent from the people you cover;
- Be willing to correct or retract your mistakes.<sup>365</sup>

Lawsuits can be expensive and traumatic even if they come to a quick end. Precautions will not prevent litigation in every case, but citizen scientists would be wise to heed experts’ advice in order to minimize the likelihood of being sued and facing liability.

## VI. CONCLUSION

The medical and scientific communities are rapidly adopting a culture of data sharing, and the expansion of open data practices is widely perceived as inevitable.<sup>366</sup> Many stakeholders are grappling with the legal and ethical implications of public access to patient-related data. For example, the prestigious Institute of Medicine is in the process of crafting a document entitled “Strategies for Responsible Sharing of Clinical Trial Data.”<sup>367</sup>

Open medical data have the potential to yield numerous benefits, including scientific discoveries, cost savings, new patient support tools, improved healthcare quality, greater government transparency, and public education.<sup>368</sup> At the same time, open data raise several complex legal and

---

364. *Practical Tips for Avoiding Liability Associated with Harms to Reputation*, DIGITAL MEDIA LAW PROJECT, <http://www.dmlp.org/legal-guide/practical-tips-avoiding-liability-associated-harms-reputation> (last updated July 22, 2008).

365. *Id.*

366. See Exec. Order No. 13,642, *supra* note 1.

367. *Activity: Strategies for Responsible Sharing of Clinical Trial Data*, INSTITUTE OF MEDICINE, <http://www.iom.edu/Activities/Research/SharingClinicalTrialData.aspx> (last visited Nov. 23, 2015) (describing the project and its timeline); see *supra* note 228 for interim report.

368. See *supra* Part III.

ethical concerns related to privacy, discrimination, erroneous research findings, and litigation.<sup>369</sup>

Scientists and policy-makers must carefully consider the varied implications of making patient-related big data available to the public. In the future, they may devise a detailed regulatory framework for citizen science.<sup>370</sup> Until then, government, industry, data custodians, and others should implement the more modest interventions proposed in this Article to protect all stakeholders: patients, researchers, businesses, and the public at large.

In his May 2013 executive order, President Obama asserted that “making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans’ lives . . . .”<sup>371</sup> Unfortunately, without well-considered responses to the legal and ethical implications of open data, the new trend may generate more harm than good. However, with careful data stewardship, society may well enjoy the new policy’s promised bounty.

---

369. *See supra* Part IV.

370. O’Connor, *supra* note 234, at 481.

371. Exec. Order No. 13,642, *supra* note 1.

# PRIVACY AND COURT RECORDS: AN EMPIRICAL STUDY

*David S. Ardia<sup>†</sup> & Anne Klinefelter<sup>††</sup>*

## ABSTRACT

As courts, libraries, and archives move to make court records available online, the increased ease of public access raises concerns about privacy. Little work has been done, however, to study how often sensitive information appears in court records and the context in which it appears. This Article fills this gap by analyzing a large corpus of briefs and appendices submitted to the North Carolina Supreme Court from 1984 to 2000. Based on a survey of privacy laws and privacy scholarship, we created a taxonomy of 140 types of sensitive information, grouped into thirteen categories. We then coded a stratified random sample of 504 court filings in order to determine the frequency of appearance of each sensitive information type and to identify relationships, patterns, and correlations between information types and various case and document characteristics.

We present several important findings. First, although a wide variety of sensitive information appears in the court records we sampled, it is not uniformly distributed throughout the records. Most of the documents contained relatively few incidences of sensitive information while a handful of documents contained a large number of pieces of sensitive information. Second, court records vary substantially in the types and frequency of sensitive information they contain. Sensitive information in seven of the categories—“Location,” “Identity,” “Criminal Proceedings,” “Health,” “Assets,” “Financial

---

DOI: <http://dx.doi.org/10.15779/Z38TR9C>

© 2015 David S. Ardia & Anne Klinefelter.

<sup>†</sup> Assistant Professor of Law, University of North Carolina School of Law, and Faculty Co-Director, UNC Center for Media Law and Policy.

<sup>††</sup> Associate Professor of Law, University of North Carolina School of Law, and Director, Kathrine R. Everett Law Library.

We could not have completed this project without the help of a dedicated team of researchers. Particular thanks is due to Esther Earbin, who oversaw the work of our coders: Dylan Arant, Nancy Brown, Alex Contarino, Kerry Dutra, Flora Feng, Emma Gilmore, Wilson Hood, Melissa Reed Muse, Haley Shell, and Josh Smith. Thank you also to Guangya (Ya) Liu for empirical research support, to Jesse Griffin and Gary Wilhelm for technical support, and to Catherine Caycedo, Daniel Parisi, Maddie Salamone, Dave Hansen, and Kate Dickson for research assistance. We are also grateful to Tamar Birckhead, John Conley, Woodrow Hartzog, Christopher Hoofnagle, William Marshall, Robert Mosteller, Helen Nissenbaum, Cathy Packer, Mary-Rose Papandrea, and the participants at the BCLT/BTLJ symposium on “Open Data: Addressing Privacy, Security, and Civil Rights Challenges” for their valuable comments and suggestions. The collection and coding of data was funded by Microsoft Corporation and the Berkeley Center for Law & Technology.

Information,” and “Civil Proceedings”—appeared much more frequently than the other categories in our taxonomy. Third, information associated with criminal proceedings, such as witness and crime victim names, is pervasive in court records, appearing in all types of cases and records. Fourth, criminal cases have disproportionately more sensitive information than civil or juvenile cases, with death penalty cases far exceeding all other case types. Fifth, appendices are generally not quantitatively different from legal briefs in terms of the frequency and types of sensitive information they contain, a finding that goes against the intuition of many privacy advocates. Sixth, there were no overarching trends in the frequency of sensitive information during the seventeen-year period we studied.

Although we found a substantial amount of sensitive information in the court records we studied, we do not take a position regarding what information, if any, courts or archivists should redact or what documents should be withheld from online access or otherwise managed for privacy protection. These largely normative questions must be answered based on a careful balancing of the competing public access and privacy interests. Nevertheless, we expect that this highly granular view of the occurrence of sensitive information in these North Carolina Supreme Court records will help policymakers and judges evaluate the potential harms to privacy interests that might arise from online access to court records. We also hope that scholars will draw on our taxonomy and empirical data to develop and ground normative arguments about the proper approach for balancing government transparency and personal privacy.

## TABLE OF CONTENTS

I.	INTRODUCTION .....	1810
II.	PUBLIC ACCESS TO COURTS AND COURT RECORDS.....	1817
A.	THE RIGHT TO ACCESS COURT PROCEEDINGS AND RECORDS .....	1817
B.	COUNTERVAILING INTERESTS .....	1824
1.	<i>Privacy and the Loss of Practical Obscurity</i> .....	1825
2.	<i>Navigating the Transition to Online Court Records</i> .....	1827
III.	A SENSITIVE INFORMATION TAXONOMY FOR COURT RECORDS .....	1828
A.	THE CHALLENGES OF CREATING A TAXONOMY OF SENSITIVE INFORMATION.....	1829
1.	<i>Building a Taxonomy on Debated Definitions of Privacy and Related Concepts</i> .....	1829
2.	<i>Charting the Piecemeal U.S. Approach to Privacy</i> .....	1832
B.	CRITERIA FOR INCLUSION IN THIS STUDY .....	1835
1.	<i>Assets</i> .....	1837
2.	<i>Civil Proceedings</i> .....	1838
3.	<i>Computer Use</i> .....	1839
4.	<i>Criminal Proceedings</i> .....	1840
5.	<i>Education</i> .....	1841
6.	<i>Employment</i> .....	1842
7.	<i>Financial Information</i> .....	1843
8.	<i>Health</i> .....	1844

2015]	PRIVACY AND COURT RECORDS	1809
	9. <i>Identity</i> .....	1845
	10. <i>Images</i> .....	1847
	11. <i>Intellectual Pursuits</i> .....	1848
	12. <i>Location</i> .....	1849
	13. <i>Sexual Activities</i> .....	1850
IV.	STUDY DESIGN AND METHODOLOGY.....	1850
A.	CORPUS OF COURT RECORDS UNDER STUDY .....	1850
B.	CODING AND ANALYSIS.....	1851
V.	RESULTS AND DISCUSSION .....	1853
A.	DESCRIPTIVE STATISTICS .....	1853
	1. <i>Sample Summary</i> .....	1853
	2. <i>Sensitive Information Summary</i> .....	1857
B.	ANALYSIS.....	1861
	1. <i>Variations Within and Among Information Categories</i> .....	1861
	2. <i>Contextual Variations</i> .....	1867
	a) <i>Case Types</i> .....	1867
	b) <i>Adults and Minors</i> .....	1870
	c) <i>Appendices</i> .....	1872
	3. <i>Temporal Variations</i> .....	1873
	4. <i>Regression Analysis</i> .....	1877
VI.	IMPLICATIONS FOR ACCESS POLICIES AND PRACTICES .....	1879
A.	IDENTIFYING WHERE PRIVACY RISKS ARE GREATEST.....	1881
	1. <i>Court Records Vary Substantially in the Sensitive Information They Contain</i> .....	1881
	2. <i>Criminal Information Is Pervasive in Court Records</i> .....	1883
	3. <i>Criminal Cases Have Disproportionately More Sensitive Information</i> .....	1884
	4. <i>Minors Deserve Additional Attention</i> .....	1886
	5. <i>It Is Unwise to Focus Exclusively on Appendices</i> .....	1887
	6. <i>Trends in Sensitive Information over Time</i> .....	1888
B.	CHALLENGES IN IMPLEMENTING PRIVACY PROTECTIVE PRACTICES .....	1889
VII.	CONCLUSION.....	1891
	APPENDIX.....	1893

## I. INTRODUCTION

Courts across the country are moving quickly to digitize their records and make them available online.<sup>1</sup> Some courts are doing this work themselves, while others are relying on third parties, such as libraries and archives, to make public access possible. All, however, are dealing with one central and unavoidable issue: privacy.<sup>2</sup>

Court records contain a variety of types of information that could be characterized as “private” or “sensitive,”<sup>3</sup> ranging from social security numbers to the names of minor children involved in sexual abuse. In *State v. Bright*, for example, a brief filed by the State of North Carolina describes the abduction and rape of a ten-year-old girl, naming the child in full on the first page and continuing to identify her by first name on nearly every subsequent page of the brief.<sup>4</sup> Similarly, in *Dean v. Cone Mills Corporation*, the plaintiff-appellant’s petition for discretionary review to the North Carolina Supreme Court includes an appendix comprising the plaintiff’s voluminous medical file and contains multiple references to his social security number, date of birth, and home address.<sup>5</sup>

Little work has been done, however, to study how often sensitive information appears in court records and the context in which it appears.<sup>6</sup>

---

1. See NATIONAL CENTER FOR STATE COURTS, TRENDS IN STATE COURTS (2014), <http://www.ncsc.org/~media/Microsites/Files/Future%20Trends%202014/2014%20NCSC%20Trends%20Report.ashx> (highlighting state courts’ efforts to move to e-filing and the conversion of paper case documents into digital images); Paul H. Anderson, *Future Trends in Public Access: Court Information, Privacy, and Technology*, in FUTURE TRENDS IN STATE COURTS 2011, at 11 (Carol R. Flango et al. eds., 2011) (reviewing the trends and issues relating to “an environment where most court systems maintain all or part of their information electronically”).

2. A wide range of technical and policy challenges continue to be a part of the effort to increase public access to court records, both in the context of electronic filing and in the digitization of older court records. But one of the central issues is privacy.

3. The terms “private” and “sensitive” in the context of personally identifiable information are not necessarily coterminous. As we discuss in Part III, there are no uniform definitions for these terms and their scope is widely debated by privacy scholars. For readability, we use “sensitive information” to refer to all types of personally identifiable information that might raise privacy concerns.

4. Brief for the State, *State v. Bright*, 505 S.E.2d 317 (N.C. Ct. App. 1998), *disc. review allowed*, 525 S.E.2d 179 (N.C. 1998), 2012 WL 6685334 (also submitted in full to the North Carolina Supreme Court).

5. Notice of Appeal and Petition for Discretionary Review, *Dean v. Cone Mills Corp.*, 322 S.E.2d 771 (N.C. 1984).

6. One important study of court records was conducted by Carl Malamud using automated software to search a large sample of court filings downloaded from the federal court’s Public Access to Court Electronic Records (PACER) system. Malamud reported to the federal courts that a significant number of social security numbers and other types

The lack of empirical data hamstrings courts and archivists who are attempting to balance privacy interests with the public's right of access, as well as scholars looking to adapt privacy law and First Amendment doctrines to deal with the flood of public records going online.

This Article helps to fill this gap in our knowledge by analyzing a large corpus of court records from the North Carolina Supreme Court. These records are held by the Kathrine R. Everett Law Library at the University of North Carolina School of Law ("UNC Law Library"), one of several libraries with copies of briefs and court filings submitted to the North Carolina Supreme Court. Through sampling and content coding of briefs and appendices filed in cases decided between 1984 and 2000, we cataloged the types of sensitive information that appeared in these records,<sup>7</sup> determined the frequency of appearance of this information, and analyzed the context in which it appeared.

As is the case for courts and archivists everywhere, UNC Law Library personnel are grappling with the question of whether—and if so, how—to limit access to sensitive information that could cause financial, reputational, or emotional harm to individuals identified in the records. Although a patchwork of court rules, statutes, and government regulations provides some guidance as to the various categories of information that might raise privacy concerns, no studies address how often this information is likely to appear in court records, in what types of documents, and the specific context of its appearance.

The results of our research are valuable for a number of reasons. First, this study provides a highly granular view of sensitive information in judicial records. Based on a survey of the laws that apply to court records

---

of sensitive information were present in the downloaded files. See John Schwartz, *An Effort to Upgrade a Court Archive System to Free and Easy*, N.Y. TIMES (Feb. 12, 2009), <http://www.nytimes.com/2009/02/13/us/13records.html> (interviewing Malamud who states that he found several types of sensitive information in the downloaded case files); see also Letter from Carl Malamud to The Honorable Lee H. Rosenthal, Chair, Committee on Rules of Practice and Procedure, Judicial Conference of the United States (Oct. 24, 2008), <https://public.resource.org/scribd/7512583.pdf> (documenting 1,669 unredacted social security numbers and other proximate sensitive information in records from 32 district courts). Computer scientist Timothy Lee followed Malamud's report with a study that found some documents submitted to courts with intended redactions were not successfully redacted. Timothy B. Lee, *Studying the Frequency of Redaction Failures in PACER*, FREEDOM TO TINKER (May 25, 2011), <https://freedom-to-tinker.com/blog/tblee/studying-frequency-redaction-failures-pacer>. Our study provides a similar, but more extensive, examination of North Carolina Supreme Court briefs and appendices.

7. We did not code for the appearance of sensitive information in the court's decisions themselves.

as well as other privacy laws and scholarship, we have identified 140 types of sensitive information that may exist in court records. By coding for these information types, we are able to determine whether their frequency of appearance is correlated with case type, document type, or other contextual factors.

Second, an understanding of the types and context of sensitive information in the North Carolina Supreme Court's case files will help policymakers and judges evaluate the potential harms to privacy interests that might arise from the disclosure of sensitive information in court briefs and related records. Although this project examines only briefs and appendices filed in the North Carolina Supreme Court during a seventeen-year period, we expect that the results of our study will be generalizable to appellate court filings in many courts throughout the United States.

Third, this research will have practical implications for court personnel and archivists as they develop rules and practices for electronic filing of court records or the digitization of older records. Based on our informal survey of archivists and law librarians around the country, we have found that digitization initiatives are proceeding without a clear or consistent strategy for addressing privacy concerns. This project will help to identify and support the development of best practices for courts and archivists developing or implementing redaction protocols or making other choices regarding access and privacy.<sup>8</sup>

Finally, this research will be valuable to privacy scholars who can use our taxonomy and data to ground their normative arguments. Legal scholars, archivists, and law librarians have written extensively about the competing interests of government transparency and personal privacy. A number of these publications focus on court records, but little research provides empirical data concerning the frequency of sensitive information in particular types of court records. Although federal court rules require that some categories of information be redacted from court filings

---

8. All providers of court records face issues with regard to quality assurance, including data entry errors and other incorrect information. Courts and other archives that redact information face significantly greater quality assurance challenges. The results of this study should help to improve accuracy and reduce the cost of implementing redaction protocols, whether they are done manually (as the majority of such protocols are handled today) or through software. See Eric O. Scott et al., *Text Mining for Quality Control of Court Records* (Mitre Corp., Case Number 14-2510), <http://mason.gmu.edu/~escott8/publications/Scott%20et%20al.,%20Text%20Mining%20for%20Quality%20Control%20of%20Court%20Records.pdf> (presented at SemADoc 2014: Semantic Analysis of Documents Workshop, 16 September 2014).

submitted via electronic filing,<sup>9</sup> and some states are following in the same direction,<sup>10</sup> debate continues about how to address transparency and privacy in the context of e-filing systems and the digitization of older court records. This project supports discussion and policy shaping in these developing areas.

The loss of “practical obscurity” lies at the heart of the debate about privacy risks from online access to court records. Although court records have historically been available to the public for review, the information in these records was for practical purposes obscure because the records were “stored in such an inaccessible fashion that only the determined and resourceful could obtain them.”<sup>11</sup> Peter Winn was one of the first to examine the loss of practical obscurity with the advent of electronic filing systems, noting that online access provided significant public benefits but raised serious privacy challenges.<sup>12</sup> Helen Nissenbaum and her coauthors

---

9. Federal e-filing rules now require redaction of several categories of data, including social security numbers and taxpayer numbers, dates of birth, names of minor children, financial account numbers, and in criminal cases, home address. *See* FED. R. APP. P. 25(a)(5); FED. R. CIV. P. 5.2; FED. R. CRIM. P. 49.1; FED. R. BANKR. P. 9037.

10. For example, since 2009 North Carolina has required e-filers to “exclude or partially describe sensitive, personal or identifying information such as any social security, employer taxpayer identification, driver’s license, state identification, passport, checking account, savings account, credit card, or debit card number, or personal identification (PIN) code or passwords from documents filed with the court. In addition, minors may be identified by initials, and, unless otherwise required by law, social security numbers may be identified by the last four numbers.” *Second Supplemental Rules of Practice and Procedure for the North Carolina eFiling Pilot Project, Rule 6.3*, N.C. COURT INFO. SYS. (Aug. 27, 2013), <https://www.efiling.nccourts.org/manual/fiCourtRules.htm> [hereinafter *N.C. eFiling Rules*] (referring to N.C.G.S. 132-1.10(d)). North Carolina also provides for non-parties to litigation to request removal or redaction of court documents available online for public viewing “if the document contains sensitive, personal or identifying information about the requester.” *Id.* at Rule 6.4 (referring to N.C.G.S. § 132-1.10(f)). An excellent review of access policies for South Dakota state court records was produced in 2005. *See* Lynn E. Sudbeck, *Placing Court Records Online: Balancing Judicial Accountability with Public Trust and Confidence*, INSTITUTE FOR COURT MANAGEMENT, COURT EXECUTIVE DEVELOPMENT PROGRAM, PHASE III PROJECT (May 2005), <https://www.ncsc.org/~media/Files/PDF/Education%20and%20Careers/CEDP%20Papers/2005/SudbeckLynnCEDPFinal32905.ashx> (recommending full access for the courts and litigants and an electronic version for the public with sensitive information redacted); *see also* D. R. Jones, *Protecting the Treasure: An Assessment of State Court Rules and Policies for Access to Online Civil Court Records*, 61 *DRAKE L. REV.* 375 (2013).

11. *Practical Obscurity*, SOC. OF AM. ARCHIVISTS, <http://www2.archivists.org/glossary/terms/p/practical-obscurity> (last visited Feb. 1, 2016); *see also* Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 *CALIF. L. REV.* 1, 4–8 (2013).

12. *See* Peter A. Winn, *Online Court Records: Balancing Judicial Accountability and Privacy in an Age of Electronic Information*, 79 *WASH. L. REV.* 307 (2004) [hereinafter Winn, *Online Court Records*]; Peter A. Winn, *Judicial Information Management in an*

also have highlighted the chasm of difference between traditional in-person public access to court records at the courthouse and Internet access through Google and electronic filing systems.<sup>13</sup> Nissenbaum also articulated the importance of protecting information-flow norms threatened by disruptive technologies and practices such as the movement to make court records electronic and widely accessible.<sup>14</sup> Others too have explored this tension of interests,<sup>15</sup> and many propose that various types of information should be protected from broad exposure.<sup>16</sup>

Archivists and law librarians also have noted the challenges posed by electronic court records and documented their own efforts to address privacy concerns while developing new projects to digitize court documents in order to expand and facilitate public access.<sup>17</sup> For example,

---

*Electronic Age: Old Standards, New Challenges*, 3 FED. CTS. L. REV. 135 (2009) [hereinafter Winn, *Judicial Information Management*].

13. Amanda Conley, Anupam Datta, Helen Nissenbaum & Divya Sharma, *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV. 772 (2012).

14. Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119 (2004) (concluding that social and practical norms regarding information flow are superior to formal bifurcations of information into categories of public or private).

15. See, e.g., Lynn M. LoPucki, *Court-System Transparency*, 94 IOWA L. REV. 481, 537 (2009) (suggesting privacy objections to highly transparent electronic federal court records should be addressed through removal of sensitive data and selective sealing of records and should not be used to shield the courts from scrutiny); Peter W. Martin, *Online Access to Court Records: From Documents to Data, Particulars to Patterns*, 53 VILL. L. REV. 855, 882–84 (2008) (warning that litigants are unlikely to adapt quickly to protect privacy, especially on behalf of non-litigants whose data may be in court filings); Will Thomas DeVries, *Protecting Privacy in the Digital Age*, 18 BERKELEY TECH. L.J. 283, 301 (2003) (“Digital technology is turning the asset of open government into a privacy nightmare. In the analog age, public records were all available, but languished in ‘practical obscurity’ in courthouse basements or isolated file cabinets.”).

16. See, e.g., Natalie Gomez-Velez, *Internet Access to Court Records*, 51 LOYOLA L. REV. 365 (2005) (recommending the use of protective orders and sealing to remove from public view high risk data elements and describing some states’ decision to exclude categories of court records from online systems); Caren Myers Morrison, *Privacy, Accountability, and the Cooperating Defendant: Towards a New Role for Internet Access to Court Records*, 62 VAND. L. REV. 921, 969–78 (2009) (recommending redaction of names of cooperating defendants and other informants while increasing transparency in the use of these law enforcement practices); Kristin A. Henderson, *Lessons from Bankruptcy Court Public Records*, 23 LEGAL REFERENCE SERVS. Q. 55, 73, 76–77 (2004) (evaluating the sensitivity of information provided in bankruptcy proceedings and supporting the American Association of Law Libraries’ proposal for redaction of sensitive information from bankruptcy court records accessible to the public through electronic case files).

17. See, e.g., Michael Whiteman, *Appellate Court Briefs on the Web: Electronic Dynamos or Legal Quagmire?* 97 L. LIBRARY J. 467, 470–77 (2005) (describing the need to preserve access to court records and discussing the challenges associated with protecting privacy, including the Northern Kentucky Law Library’s decision to refrain

the Montana State Law Library, which began scanning and posting Montana Supreme Court opinions and briefs in 1996, removed records it had already posted online to extract exhibits and appendices because the library found that these records contained a variety of sensitive information; the library ultimately reposted only the briefs with redactions.<sup>18</sup>

The goal of this study is to inform these scholarly and policy discussions about the appropriate balance between public access and privacy in the context of court records. We begin in Part II by noting that court records present a special challenge for privacy advocates. Unlike in many other areas of privacy law, court records are presumptively open to the public. Part II describes the origin of the right of public access to court records and examines its scope under the federal Constitution, common law, statutory law, and court rules. As we note in Part II, not all repositories of court records are obligated by law to provide public access. For many librarians and archivists, the question is not what the law requires, but rather what is the best approach for ensuring the protection of privacy interests while at the same time informing the public about the functioning of the court system.

In Part III, we survey privacy laws and privacy scholarship to create a taxonomy of sensitive information in court records.<sup>19</sup> Based on this survey,

---

from scanning appendices to briefs filed in the Kentucky Supreme Court). Archivists have been dealing with privacy concerns in a broad range of materials for many years and are considering how born-digital and digitized materials can be managed to address access and privacy. *See, e.g.*, Christopher A. Lee & Kam Woods, *Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage*, in PROCEEDINGS OF THE MEMORY OF THE WORLD IN THE DIGITAL AGE: DIGITIZATION AND PRESERVATION: AN INTERNATIONAL CONFERENCE ON PERMANENT ACCESS TO DIGITAL DOCUMENTARY HERITAGE (2012), <http://ils.unc.edu/caltee/p298-lee.pdf>.

18. *See* Tammy A. Hinderman, *State Law Library Gets a 21st Century Makeover*, 32 MONT. L. REV. 6 (2007). According to Montana State Law Library reference librarian Tammy Hinderman, the library began providing online access to the Montana Supreme Court records in 2006, before realizing the records contained sensitive information that could facilitate identity theft and other privacy harms. *Id.* at 7. She explains that the library then removed all exhibits and appendices from the electronic version of the documents and redacted some information from the briefs before reposting them to the Internet. *Id.* In Kentucky, the Chase College of Law Library of Northern Kentucky University began its scanning of briefs from the state's supreme court by omitting appendices, both to address privacy concerns and to limit the burden on the library. *See* Whiteman, *supra* note 17, at 477.

19. Taxonomies of sensitive information appear throughout the law of the United States and other jurisdictions, and electronic filing systems at both the federal and state

we identified 140 types of sensitive information that might appear in court records and grouped them into thirteen categories. Part III explains the justifications—and shortcomings<sup>20</sup>—of our taxonomic approach and describes the sensitive information types we coded for in this project. Of course, not everyone will agree with our taxonomy. That is to be expected, given that privacy is itself a contested concept. Nevertheless, the taxonomy has proven to be helpful to us in the identification of the privacy risks that can arise from the public disclosure of court records at a time when privacy laws have limited or unclear application to such records. Moreover, we think our extensive taxonomy will be useful to others who wish to understand the broad range of privacy interests implicated by public records.

In Part IV, we provide an overview of our study design and methods. In short, we analyzed a stratified random sample of 504 court documents pulled from the briefs and other filings submitted to the North Carolina Supreme Court from 1984 to 2000. After performing content coding of the documents, we determined the frequency of appearance of each sensitive information type and identified relationships, patterns, and correlations between different information types and other coded variables, including trends over time.

In Part V, we present a summary of our findings. We begin by providing descriptive statistical information about the court records in our sample and the sensitive information they contain. We then examine the extent to which different types of sensitive information are related to various case and document characteristics. Although we suggest ways in which our data can aid in the assessment of the privacy risks that might arise from public access to court records, it is not our aim to tell courts or archivists what information, if any, should be redacted or what documents should be withheld from online access or otherwise managed for privacy protection.<sup>21</sup>

Instead, in Part VI, we discuss how our study can inform the debate about privacy and court records and how our results can help to identify and remedy some of the challenges courts and archivists are likely to face if they decide to implement procedures for addressing privacy concerns in

---

level rely extensively on predefined lists of information types that must be handled with special care. *See infra* notes 89–98 and accompanying text.

20. Scholars have long criticized this approach because it relies on debated definitions of privacy, ignores contextual variations in privacy, and presents implementation challenges because of these definitional and contextual problems. We discuss these concerns and how we dealt with them in Part III.

21. We plan to address these normative questions in subsequent articles.

court records. We made several important findings in this regard. First, although a wide variety of sensitive information appears in the court records we sampled, it is not uniformly distributed throughout the records. Most of the documents contained relatively few incidences of sensitive information while a handful of documents contained a large number of pieces of sensitive information. Second, we found that court records vary substantially in the types and frequency of sensitive information they contain. Sensitive information in seven categories—“Location,” “Identity,” “Criminal Proceedings,” “Health,” “Assets,” “Financial Information,” and “Civil Proceedings”—appeared much more frequently than information in the other categories we identified. Third, we found that information associated with criminal proceedings, such as witness and crime victim names, is pervasive in court records, appearing in all types of cases and records. Information in the “Criminal Proceedings” category not only appeared in most of the documents we reviewed, but also appeared more often in those documents than any other category of sensitive information. Fourth, the data showed that criminal cases have disproportionately more sensitive information than civil or juvenile cases., with death penalty cases far exceeding all other case types. Fifth, we found that appendices are generally not quantitatively different than legal briefs in terms of the frequency and types of sensitive information they contain, a finding that goes against the intuition of many privacy advocates. Sixth, we saw no overarching trends in the frequency of sensitive information during the seventeen-year period under study.

We close by providing some suggestions for courts and archivists seeking to manage sensitive information in court records. A number of practices have been introduced or recommended, including redaction of electronic records, redaction of both electronic and print records, removal of categories of court records from Internet access, and increased filing of court documents under seal. Our research will help courts and archivists evaluate these approaches.

## II. PUBLIC ACCESS TO COURTS AND COURT RECORDS

### A. THE RIGHT TO ACCESS COURT PROCEEDINGS AND RECORDS

Public access to the courts has a long and venerated history in America, even predating enactment of the United States Constitution.<sup>22</sup>

---

22. *See, e.g.,* *Leucadia, Inc. v. Applied Extrusion Techs., Inc.*, 998 F.2d 157, 161 (3d Cir. 1993) (concluding that “[t]he existence of this right, which antedates the Constitution and which is applicable in both criminal and civil cases, is now ‘beyond

This openness serves many salutary functions, including ensuring that our system of justice functions fairly and is accountable to the public.<sup>23</sup> As Chief Justice Warren Burger noted in *Richmond Newspapers, Inc. v. Virginia*:

The early history of open trials in part reflects the widespread acknowledgment, long before there were behavioral scientists, that public trials had significant community therapeutic value. Even without such experts to frame the concept in words, people sensed from experience and observation that, especially in the administration of criminal justice, the means used to achieve justice must have the support derived from public acceptance of both the process and its results.<sup>24</sup>

Public access also has been extended to many of the records associated with court proceedings.<sup>25</sup> Access to judicial records plays a critical role in fostering public awareness about the operation of the courts because so few people are able to attend court proceedings in person and because most courts do not generally allow live or archival recordings. The movement by courts and archivists to allow online access to court records has made it possible for many more people to stay informed about the functioning of the judicial system.<sup>26</sup> Online access also has a leveraging effect because it makes it possible for the media to cover court proceedings at a lower cost and allows for greater depth of reporting at a time when many media

---

dispute.”); Winn, *Online Court Records*, *supra* note 12, at 307 (noting that “the legal system has inherited from the Enlightenment a presumption of openness”); Conley et al., *supra* note 13, at 785 (observing that “the right to open courts and their records is actually as longstanding as our right to the courts and to justice itself”).

23. See *Globe Newspaper Co. v. Super. Ct.*, 457 U.S. 596, 606 (1982) (“Public scrutiny of a criminal trial enhances the quality and safeguards the integrity of the factfinding process, with benefits to both the defendant and to society as a whole.”).

24. *Richmond Newspapers, Inc. v. Virginia*, 448 U.S. 555, 570–71 (1980).

25. The public’s right of access to court records is not absolute and may be restricted in some circumstances. See *infra* notes 30–33 and accompanying text.

26. See, e.g., Lynn E. Sudbeck, *Placing Court Records Online: Balancing Judicial Accountability with Public Trust and Confidence: An Analysis of State Court Electronic Access Policies and a Proposal for South Dakota Court Records*, 51 S.D. L. REV. 81, 91 (2006) (noting that a “frequently mentioned benefit” of electronic access to court records is that it responds “to the needs of South Dakota’s rural court users, that is, [it] ‘levels the geographic playing field’ by allowing persons located in great distances from the courthouse to access public information” (citation omitted)). Online access to court records also allows litigants, lawyers, and educators to scrutinize legal strategy and rhetoric. See Anna P. Hemingway, *Making Effective Use of Practitioners’ Briefs in the Law School Curriculum*, 22 ST. THOMAS L. REV. 417 (2010) (advocating use of practitioners’ briefs to teach persuasive writing and legal analysis).

organizations are cutting back on the number of reporters assigned full-time to the courts.<sup>27</sup>

Although public access to court records is longstanding and deeply ingrained in our legal system, courts can—and often do—impose limits on public access. The First Amendment provides a right of access to court proceedings and to many records,<sup>28</sup> as does federal and state common law.<sup>29</sup> These rights, however, are not absolute.<sup>30</sup> While the precise standard that a court must apply will vary depending on the source of the public’s right of access, in general courts must at least conclude that the interest in prohibiting disclosure outweighs the strong presumption of public access. In *Nixon v. Warner Communications*, for example, the United States Supreme Court instructed that “[e]very court has supervisory power over its own records and files” and that the federal

---

27. See *Panel One: General Discussion on Privacy and Public Access to Court Files*, 79 *FORDHAM L. REV.* 1, 13 (2010) (quoting testimony of Lucy Dalglish before the Privacy Subcommittee of the Judicial Conference Standing Committee on the Federal Rules).

28. See, e.g., *Richmond Newspapers*, 448 U.S. at 575 (finding First Amendment right of public access to criminal trials and noting that “[i]n guaranteeing freedoms such as those of speech and press, the First Amendment can be read as protecting the right of everyone to attend trials so as to give meaning to those explicit guarantees”). Although the U.S. Supreme Court has not explicitly held that a First Amendment right of access applies in civil cases, most of the federal circuits that have addressed this issue have recognized such a right. See, e.g., *Westmoreland v. Columbia Broad. Sys., Inc.*, 752 F.2d 16, 23 (2d Cir. 1984); *Publicker Indus., Inc. v. Cohen*, 733 F.2d 1059, 1067–71 (3d Cir. 1984). Courts have also applied a constitutional right of access to the judicial records associated with criminal and civil proceedings. See, e.g., *Associated Press v. United States Dist. Court. for Cent. Dist. of Cal.*, 705 F.2d 1143, 1145 (9th Cir. 1983); *In re Search Warrant*, 855 F.2d 569, 573 (8th Cir. 1988); *Newsday LLC v. Cnty. of Nassau*, 730 F.3d 156, 164 (2d Cir. 2013); *Publicker Indus.*, 733 F.2d at 1074. *But see Zenith Radio Corp. v. Matsushita Elec. Indus. Co.*, 529 F. Supp. 866, 908 (E.D. Pa. 1981) (“With respect to the question whether the common law right to inspect and copy [discovery materials] has a constitutional dimension, we conclude that it does not.”).

29. See *Nixon v. Warner Commc’ns*, 435 U.S. 589, 597 (1978) (recognizing a federal common law right to “inspect and copy public records and documents, including judicial records and documents”); Richard J. Peltz et al., *The Arkansas Proposal on Access to Court Records: Upgrading the Common Law with Electronic Freedom of Information Norms*, 59 *ARK. L. REV.* 555, 591–94 (2006) (discussing various state approaches).

30. When the right of public access arises under the First Amendment, “it must be shown that the denial [of access] is necessitated by a compelling governmental interest, and is narrowly tailored to serve that interest.” *Globe Newspaper Co. v. Super. Ct.*, 457 U.S. 596, 607 (1982). When the right of access is merely a common right, courts have more leeway in denying access and can balance the presumption of public access against other interests, including the possibility of prejudicial pretrial publicity; the danger of impairing law enforcement or judicial efficiency; and the protection of the legitimate privacy interests of litigants and other trial participants, such as witnesses, victims, and jurors. See, e.g., *Nixon*, 435 U.S. at 598.

common law right of access could be denied when “court files might . . . become a vehicle for improper purposes.”<sup>31</sup> Among the improper purposes the Court noted were uses “to gratify private spite or promote public scandal,” as “reservoirs of libelous statements for press consumption,” and as a source of unfair competitive “business information.”<sup>32</sup> Although there is considerable variation among the states, state common law rights of access also are typically qualified rights, allowing courts to restrict public access if an overriding interest supports closure or sealing of specific information. In California, for example, court records are “presumptively open to the public and [court proceedings and records] should not be closed except for compelling countervailing reasons.”<sup>33</sup>

In addition to constitutional and common law rights of access, a number of state and federal statutes also provide a public right of access to court records. At the federal level, access to court records is governed by rules and policies promulgated by the Administrative Office of the U.S. Courts on behalf of the federal judiciary pursuant to the Rules Enabling Act.<sup>34</sup> At the state level, a variety of statutory authority provides for and impacts public access to judicial records. For example, every state has a public records statute, although not all of these statutes explicitly address access to court records.<sup>35</sup> In those states that do have a public records law that covers judicial records, rights of access are typically governed by both the statute and court rules.<sup>36</sup>

---

31. *Nixon*, 435 U.S. at 598.

32. *Id.*

33. *Pantos v. City & Cnty. of S.F.*, 198 Cal. Rptr. 489, 492 (Cal. Ct. App. 1984) (citations omitted).

34. 28 U.S.C. §§ 2071–2077. The Rules Enabling Act authorizes the Supreme Court to prescribe general rules of practice and procedure and rules of evidence for the federal courts. Pursuant to Section 2073 of the Rules Enabling Act, the U.S. Judicial Conference has established procedures to govern the work of its Standing Committee and its advisory rules committees. *See* UNITED STATES COURTS, CRIMINAL JUSTICE ACT GUIDELINES, GUIDE TO JUDICIARY POLICY § 440 (2014), <http://www.uscourts.gov/file/2932/download>.

35. *See* REPORTERS COMMITTEE FOR FREEDOM OF THE PRESS, OPEN GOVERNMENT GUIDE (5th ed. 2006) (providing summaries of public records laws in all U.S. jurisdictions), <http://www.rcfp.org/ogg/index.php>. When a state’s public records law is silent, the state’s highest court may define the scope and procedures for public access.

36. *See, e.g., Anderson v. Home Ins. Co.*, 924 P.2d 1123, 1126 (Colo. App. 1996) (interpreting the state’s public records law and court rules to hold that there is a strong presumption that all court records are open); *Doe v. New York Univ.*, 786 N.Y.S.2d 892, 899 (N.Y. Sup. Ct. 2004) (applying both statutory law and court rules).

As a result, many states have multiple overlapping sources of law that require—and potentially limit—public access.<sup>37</sup> In North Carolina, which is illustrative of the law in many states, public access is governed by, *inter alia*, a common law right of access,<sup>38</sup> a constitutional right of access rooted in both the First Amendment to the U.S. Constitution and article 1, section 18 of the N.C. Constitution, which states that “[a]ll courts shall be open,”<sup>39</sup> and court rules that specify how court records are to be handled, including rules for electronic-filing.<sup>40</sup>

Furthermore, the North Carolina General Assembly, through the state’s public records law (“NC PRL”) and other statutes, has both expanded and narrowed the public’s right of access.<sup>41</sup> The NC PRL, which states that all state records “are the property of the people,” is applicable to every agency of the North Carolina government, including the judiciary.<sup>42</sup>

---

37. See Richard J. Peltz, et al., *The Arkansas Proposal on Access to Court Records: Upgrading the Common Law with Electronic Freedom of Information Norms*, 59 ARK. L. REV. 555, 591 (2006) (noting that “[s]ome states decided that only one type of law was necessary to adequately provide a right of access, while others applied multiple types of law to provide more depth to their access law”).

38. See *Virmani v. Presbyterian Health Servs. Corp.*, 515 S.E.2d 675, 691 (N.C. 1999) (observing that “[a]t least since 1887, this Court has recognized a common law right of the public to inspect public records”). As with the federal common law, the common law right of access in North Carolina is a qualified right. The decision to deny access “is left to the sound discretion of the trial courts, a discretion to be exercised in light of the relevant facts and circumstances of the particular case.” *In re Investigation into Death of Cooper*, 683 S.E.2d 418, 425 (N.C. App. 2009) (internal quotation marks omitted).

39. *Virmani*, 515 S.E.2d at 692 (holding that the N.C. Constitution guarantees a qualified constitutional right on the part of the public to attend civil court proceedings and access court records). In the words of the North Carolina Supreme Court: “That courts are open is one of the sources of their greatest strength.” *Raper v. Berrier*, 97 S.E.2d 782, 784 (N.C. 1957).

40. For example, N.C.’s eFiling Rule 6.3 states, in part:

Except where otherwise expressly required by law, filers must comply with G.S. 132-1.10(d) to exclude or partially describe sensitive, personal or identifying information such as any social security, employer taxpayer identification, driver’s license, state identification, passport, checking account, savings account, credit card, or debit card number, or personal identification (PIN) code or passwords from documents filed with the court.

*N.C. eFiling Rules*, *supra* note 10, at Rule 6.3.

41. See *In re Investigation into Death of Cooper*, 683 S.E.2d at 425; N.C. GEN. STAT. § 132-1 (2015).

42. N.C. GEN. STAT. § 132-1(b) (2015) (“The public records and public information compiled by the agencies of North Carolina government or its subdivisions are the property of the people. Therefore, it is the policy of this State that the people may

The NC PRL does not grant court records any special dispensation from public access requirements, except to define two narrow exceptions to the law. The first exception allows for the withholding of settlements in medical malpractice actions against public hospitals.<sup>43</sup> The other exception makes arrest and search warrants confidential until they have been returned.<sup>44</sup> Various statutes outside the NC PRL also treat some court documents as confidential. These include records of grand jury proceedings;<sup>45</sup> most adoption records;<sup>46</sup> and reports of cases of juvenile abuse, neglect, or dependency.<sup>47</sup> Other than these significant exceptions, almost all court records are subject to public inspection under the NC PRL unless otherwise specifically restricted by law.<sup>48</sup>

Given this overlapping and sometimes ambiguous legal authority, it should come as no surprise that individual judges and court clerks frequently struggle with how to implement the public's right of access to court records. As Amanda Conley, Anupam Datta, Helen Nissenbaum, and Divya Sharma note, "restrictions on access trickle down from state and federal appellate courts to the local courthouses themselves, where state and local law, custom, and in some cases simply the whims of court clerks determine which information in the court record will actually be made available to the public, and how."<sup>49</sup>

Moreover, librarians and archivists, who may not be bound by law to provide public access to court records,<sup>50</sup> have an even broader range of

---

obtain copies of their public records and public information free or at minimal cost unless otherwise specifically provided by law.”)

43. N.C. GEN. STAT. § 132-1.3(a) (2015).

44. N.C. GEN. STAT. § 132-1.4(k) (2015).

45. N.C. GEN. STAT. § 15A-623 (2015).

46. N.C. GEN. STAT. § 48-9-102 (2015).

47. N.C. GEN. STAT. § 7B-2901 (2015).

48. *See News & Observer Pub. Co. v. Poole*, 412 S.E.2d 7, 19 (N.C. 1992) (“[W]e hold that in the absence of clear statutory exemption or exception, documents falling within the definition of ‘public records’ in the Public Records Act must be made available for public inspection.”).

49. Conley et al., *supra* note 13, at 787.

50. The applicability of public records statutes to publicly supported libraries' collections is not well established. Public libraries have been described as requiring autonomy to add and withdraw materials from their collections, at least in the context of First Amendment analysis. *See United States v. Am. Library Ass'n*, 539 U.S. 194, 195 (2003) (Rehnquist, C.J., plurality opinion) (“To fulfill their traditional missions of facilitating learning and cultural enrichment, public libraries must have broad discretion to decide what material to provide to their patrons.”). State archives, however, tend to have statutory requirements for providing access to public records. *See Carol D. Billings, State Government Efforts to Preserve Electronic Legal Information*, 96 L. LIBRARY J. 625, 626 (2004) (noting that “most state libraries that operate the depository programs and

options for dealing with sensitive information in court records. For many, the question is not what the law requires, but rather what is the best policy for ensuring the protection of privacy interests while at the same time informing the public about the functioning of the court system.<sup>51</sup> As a result, some libraries exclude whole categories of records from public access,<sup>52</sup> whereas others engage in targeted redactions of sensitive information based either on their own assessment of what is private<sup>53</sup> or on the electronic filing rules adopted by their courts.<sup>54</sup> Alternatively, several libraries have adopted a middle-ground approach. They provide mediated access to court records, allowing only bibliographic information to be discoverable on the Internet, not the contents of the records themselves,<sup>55</sup> or limiting access to unaltered briefs to registered library users.<sup>56</sup>

---

state archives with responsibility for preserving records lack rule-making and enforcement powers to require compliance”).

51. *See, e.g.*, Hinderman, *supra* note 18, at 7 (discussing the Montana State Law Library’s efforts to balance privacy and public access concerns); Whiteman, *supra* note 17, at 477 (describing the approach taken by Northern Kentucky University’s law library). Even if a library or other archive decides to make case files available without any restrictions on access, it should not face legal liability if the records contain information that violates privacy law. *See Cox Broad. Corp. v. Cohn*, 420 U.S. 469, 496 (1975) (“Once true information is disclosed in public court documents open to public inspection, the press cannot be sanctioned for publishing it.”).

52. *See* Hinderman, *supra* note 18, at 7 (describing the exclusion of appendices); Whiteman, *supra* note 17, at 477 (same).

53. The Montana State Law Library redacted social security numbers, dates of birth and other “obviously private information” from the briefs in its database of supreme court briefs. Hinderman, *supra* note 18, at 7. The Blakley Law Library at the Arizona State University Sandra Day O’Connor College of Law has posted digital versions of state appellate and supreme court briefs to the Internet with the caveat, “Certain types of personal information may have been removed from briefs on the Arizona Memory Project to allow for online publication.” *About Collection*, ARIZONA MEMORY PROJECT, <http://azmemory.azlibrary.gov/cdm/landingpage/collection/asuross> (last visited May 31, 2015).

54. *See* Faye Jones & Caroline Osborne, *Lessons Learned: Creating Digital Collections and Privacy: Best Practices*, Presentation at the Southeastern Association of Law Libraries Annual Meeting (April 16, 2015) (presentation slides on file with authors) (comments of Faye Jones, describing the Florida State University College of Law Library’s collaboration with other Florida law libraries to provide Internet access to state supreme court briefs and citing Florida public records laws (FLA. STAT. §§ 119.01–.15 as well as FLA. R. JUD. ADM. 2.420), which outline confidentiality guidelines for filing of court records including redaction).

55. *Id.* (comments of Caroline Osborne, explaining Washington and Lee Law Library’s project to digitize and not redact copies of Virginia Supreme Court briefs, to store the digital briefs in a “dark archive,” and to develop policies and procedures for responding to requests for individual briefs, citing state statutes on freedom of information (VA. CODE §§ 2.2-3700–3714), prohibition of posting certain information

## B. COUNTERVAILING INTERESTS

Court records contain a variety of information that can cause harm to individual, organizational, and governmental interests. A court's file for a single case may consist of thousands of documents, including motions, pleadings, briefs, transcripts, exhibits entered into evidence, and records and responses produced during pre-trial discovery that have been filed with the court.<sup>57</sup> For individuals, information ranging from social security numbers to sexual history can appear in these documents raising, among other concerns, the risk of identity theft and reputational harm.<sup>58</sup> For businesses and other organizations, court records can contain trade secrets and other confidential information.<sup>59</sup> For the government, information in court records such as the names of confidential informants and descriptions of intelligence gathering techniques can potentially harm national security or undermine law enforcement efforts.<sup>60</sup> Although all of these countervailing interests are worthy of study, our focus is on the impact that the disclosure of sensitive information in court records can have on individuals.

Given that "[t]he courts are a stage where many of life's dramas are performed, where people may be shamed, vindicated, compensated, punished, judged, or exposed,"<sup>61</sup> it is natural that court records, which

---

to the Internet (VA. CODE § 17.1-293), and personal information privacy (VA. CODE § 59.1.443.2)).

56. See *Policies for Utah Court Briefs*, HOWARD W. HUNTER LAW LIBRARY, J. RUBEN CLARK LAW SCHOOL, BRIGHAM YOUNG UNIVERSITY, [http://digitalcommons.byu.edu/utah\\_court\\_briefs/policies.html](http://digitalcommons.byu.edu/utah_court_briefs/policies.html) (last visited Feb. 1, 2016) (explaining that briefs are "supplied to the Hunter Law Library by the courts for the purposes of legal scholarship and academic research. The Hunter Law Library provides this collection as authorized by the Utah Courts. The Law Library is not responsible for the selection or content of individual records.").

57. See Conley et al., *supra* note 13, at 781 (noting that "[e]ach and every form filled out by the parties, their lawyers, or by related third parties (witnesses, jurors, etc.) potentially contains vast amounts of personal data including home or school addresses, places of employment, birthdates, and, in many cases, Social Security numbers."). Some documents such as sealed discovery materials, see *Leucadia, Inc. v. Applied Extrusion Techs, Inc.*, 998 F.2d 157, 163–65 (3d Cir. 1993), and certain financial information about the parties, see *United States v. Lexin*, 434 F. Supp. 2d 836, 849 (S.D. Cal. 2006), are often excluded from the public court record.

58. See *infra* Part III.

59. See Kyle J. Mendenhall, *Can You Keep A Secret? The Court's Role in Protecting Trade Secrets and Other Confidential Business Information from Disclosure in Litigation*, 62 DRAKE L. REV. 885 (2014).

60. See Laura K. Donohue, *The Shadow of State Secrets*, 159 U. PA. L. REV. 77, 78 (2010).

61. Conley et al., *supra* note 13, at 774.

serve as a chronicle of these dramas, are littered with private and sensitive information. In fact, they are full of information not just about the parties in a case, but also about witnesses, family members, victims, and jurors, among other individuals who are brought willingly or unwillingly into a legal dispute.

Although concerns about private information in court records existed long before the Internet, many commentators see the move to electronic court records as effectuating a qualitative shift in the balance between the competing interests of public access and individual privacy. Not so long ago it was difficult and time-consuming to access and search an entire case file. Today, with the advent of electronic court records and online access, it takes little effort to find and link information across cases, courts, and states. The following sections highlight the most pressing concerns that arise from the transition to online court records. We then dive much more deeply into these issues in Parts III and VI.

### 1. *Privacy and the Loss of Practical Obscurity*

Courts, like other institutions in our society, are in the midst of a transformation. The largely paper-based world of the twentieth century is giving way to an interconnected, electronic world where physical and temporal barriers to public access are evaporating. Over the past decade, courts across the country have been moving with alacrity to digitize their records and make them available to the public online.<sup>62</sup> Some courts are doing this work themselves, while others are relying on third parties, such as libraries and other archives, to make online access to historical records possible. A growing number of courts also require litigants to file their pleadings, motions, and other documents in electronic format.<sup>63</sup>

---

62. See John T. Matthias, *E-Filing Expansion in State, Local, and Federal Courts 2007*, in *FUTURE TRENDS IN STATE COURTS 2007*, at 34 (highlighting state courts' efforts to move to e-filing and the conversion of paper case documents into digital images), <http://ncsc.contentdm.oclc.org/cdm/ref/collection/tech/id/570>; HON. PAUL H. ANDERSON, *FUTURE TRENDS IN PUBLIC ACCESS: COURT INFORMATION, PRIVACY, AND TECHNOLOGY* 11 (2011) (reviewing the trends and issues relating to "an environment where most court systems maintain all or part of their information electronically").

63. See, e.g., Peter W. Martin, *Online Access to Court Records—from Documents to Data, Particulars to Patterns*, 53 *VILL. L. REV.* 855, 872 (2008) ("By the end of 2007, electronic filing was an option in nearly all federal trial courts and was mandatory in a large number."); Eric J. Magnuson & Samuel A. Thumma, *Prospects and Problems Associated with Technological Change in Appellate Courts: Envisioning the Appeal of the Future*, 15 *J. APP. PRAC. & PROCESS* 111, 114 (2014) ("By late 2012, all federal courts of appeals were using electronic filing (e-filing)."); Matthias, *supra* note 62, at 34 (reporting

As discussed in the previous section, court records have for centuries been open for public review. Yet the difficulty of actually accessing individual records—for example, traveling to the courthouse, identifying the relevant case, finding the sought after document, and copying the information—made the information in these records practically obscure in the sense that private and sensitive information could remain in the records without creating a significant risk of harm. Today, this practical obscurity is vanishing. Although the specifics of electronic access vary by state (and sometimes by court), in most federal courts and many state jurisdictions anyone can access a court’s electronic case database through a website interface.<sup>64</sup> That interface typically provides the ability to search by party names, case type, keywords, and other information, as well as providing case-by-case browsing. If users wish to copy a document, they can usually do so by downloading it as a PDF file.<sup>65</sup>

The loss of practical obscurity that has resulted from this nearly frictionless access to court records lies at the heart of the debate about the privacy risks arising from online access. The Supreme Court recognized the importance of practical obscurity in holding that rap sheets aggregating public—but difficult to assemble—information qualify for a privacy exemption from disclosure under the federal Freedom of Information Act.<sup>66</sup> The Court stated, “Plainly there is a vast difference between the public records that might be found after a diligent search of

---

that as of 2007, twenty-six states had adopted court rules enabling e-filing statewide or in at least one court).

64. Some courts charge for access, some merely require registration, while others do not require either payment or registration.

65. In jurisdictions that have public records laws that cover court records, a requester may even be entitled to a copy of a court’s entire case database, though some limitations might apply. *See LexisNexis Risk Data Mgmt. Inc. v. N.C. Admin. Office of Courts*, 776 S.E.2d 651, 652 (N.C. 2015) (finding that the court’s Automated Criminal/Infraction System (ACIS) database was a public record under the North Carolina Public Records Act subject to a limiting statutory provision requiring requesters to secure a nonexclusive contract and pay for reasonable cost recovery).

66. *U.S. Dept. of Justice v. Reporters Committee for Freedom of the Press*, 489 U.S. 749, 762–63 (1989) (describing the view that aggregated information provides no more privacy harm than its discrete components as a “cramped notion of personal privacy”). It should be noted that the Court’s decision in *Reporters Committee for Freedom of the Press* did not address access to court records, but rather a request for access under FOIA to a database of criminal history information compiled by the FBI. *Id.* at 751–52. The standard for determining whether public access can be denied under FOIA is less demanding than the standard for restricting access to court records; all that the government was required to show was that disclosure “could reasonably be expected to constitute an unwarranted invasion of personal privacy.” *Id.* at 756 (quoting 5 U.S.C. § 552(b)(7)(C)).

courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information.”<sup>67</sup>

## 2. *Navigating the Transition to Online Court Records*

As a result of these and other concerns, court administrators, judges, lawyers, librarians, and legislators are in active discussion about how to navigate the transition to online court records.<sup>68</sup> Privacy scholars have also been trying to influence this transition. Indeed, a number of legal scholars consider practical obscurity to be a stand-in for privacy interests and now, with the loss of this obscurity, are suggesting that courts and archivists should implement various approaches to obscuring sensitive information in court records.<sup>69</sup> Other scholars also have explored the tension between privacy and public access to court records, with some recommending a substantial curtailment of public access through redaction of electronic and print records, restricted public access, removal of categories of court records from Internet access, and increased filing of court documents under seal.<sup>70</sup>

Although important theoretical work is being done with regard to the nature and extent of the privacy interests implicated by public access to court records,<sup>71</sup> we are only just beginning to develop a sufficient body of

---

67. *Id.* at 764.

68. See Conley et al., *supra* note 13, at 776 (noting that “public and internal deliberations over state access policies have remained actively in progress”).

69. See Hartzog & Stutzman, *supra* note 11, at 41 (“Obscurity obligations would not aim to completely curtail information disclosure; rather, they would seek to minimize the likelihood of discovery, comprehension, or contextualization.”); Steven C. Bennett *Pleadings, Privacy and Ethics: Protecting Privacy in Litigation Documents*, 2 REYNOLDS CT. & MEDIA L.J. 25 (2012); Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137 (2002); Will T. DeVries, *supra* note 15.

70. See, e.g., Gomez-Velez, *supra* note 16, at 431–32 (examining the decisions of some states to exclude categories of court records from online systems); Morrison, *supra* note 16, at 925–27 (recommending redaction of identifying information of cooperating defendants and other informants while increasing transparency in using these law enforcement practices); Henderson, *supra* note 16, at 76–77 (supporting the American Association of Law Libraries’ advocacy for redaction of sensitive information from bankruptcy court records accessible to the public through electronic case files). Not all scholars argue for restricting public access. See, e.g., Lynn LoPucki, *The Politics of Research Access to Court Data*, 80 TEX. L. REV. 2161 (2002) (arguing against selective restriction of access to court records to enable better empirical research about the courts).

71. See, e.g., Nissenbaum, *supra* note 14, at 136–38 (concluding that accepted social and practical norms for information flows are superior to formal bifurcations of information into categories of public or private); Hartzog & Stutzman, *supra* note 11, at 3–4 (suggesting that the concept of “online obscurity” is a critical component of online

research that examines the risks to privacy when court records are made available through the Internet compared with long-standing public access that was practically obscure due to the logistical barriers to access.

Our present research helps to fill this gap in our knowledge. Empirical data about the frequency and context of sensitive information in the North Carolina Supreme Court's files will allow policymakers and scholars to better understand and evaluate the range of privacy risks that can arise from online court records.

### III. A SENSITIVE INFORMATION TAXONOMY FOR COURT RECORDS

Our project draws on the longstanding approach to privacy of identifying certain types of information that present risks of harm that can be reduced through restrictions on public exposure. Taxonomies of sensitive information appear throughout the law of the United States and other jurisdictions, and electronic filing systems at both the federal and state level rely extensively on predefined lists of information types that must be handled with special care.<sup>72</sup> The use of sensitive information taxonomies is pervasive because they provide an attractive, seemingly simple solution, for balancing privacy and competing interests. Indeed, this approach to privacy is the basis of much of privacy law.<sup>73</sup>

---

privacy and developing an analytical framework for use by lawmakers and courts); Solove, *supra* note 69, at 1176–78 (criticizing the “secrecy paradigm” in privacy discourse and suggesting that there is an “expectation of limits on the degree of accessibility” to public records).

72. *See, e.g.*, HIPAA Privacy Rule, 45 C.F.R. § 164.514(b)(2) (2015) (listing seventeen specific identifiers to be removed to “de-identify” personal health information); Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of such Data, art. 2, sec. (a), 1995 O.J. (L 281) 31, 38, <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:31995L0046&from=en> (defining personal data). As Paul Ohm has noted, although many scholars have turned away from the list-based approach to privacy protection, U.S. law remains largely grounded in this model. Paul Ohm, *Sensitive Information*, 88 S. CAL. L. REV. 1125, 1125–32 (2015).

73. *See* Ohm, *supra* note 72, at 1128–29 (“The great variety of regulations, law, technical standards, and corporate practices that have been implemented to protect the privacy of information stored in databases share at their core this unifying construct [of sensitive information.]”).

A. THE CHALLENGES OF CREATING A TAXONOMY OF SENSITIVE INFORMATION

1. *Building a Taxonomy on Debated Definitions of Privacy and Related Concepts*

The creation of a comprehensive taxonomy of sensitive information types is a challenge because privacy law and policy are not grounded in a coherent understanding of or approach to privacy.<sup>74</sup> Moreover, some conceptions of privacy simply do not lend themselves to a sensitive-information approach. In addition, disagreement about the proper role of related concepts of confidentiality, practical obscurity, and contextual privacy increases the difficulty of creating a taxonomy of sensitive information.

One of the core problems is the lack of consensus about the underlying interests and risks that define privacy. Financial integrity,<sup>75</sup> personal safety,<sup>76</sup> non-discrimination,<sup>77</sup> confidential access to professional advice,<sup>78</sup>

---

74. Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477, 477–78 (2006) (“Privacy is a concept in disarray. Nobody can articulate what it means. As one commentator has observed, privacy suffers from ‘an embarrassment of meanings.’”).

75. Identity-theft statutes prohibit the disclosure of data such as financial account numbers and PIN codes. *See, e.g.*, N.C. GEN. STAT. § 14-113.20(b) (2015); MASS. GEN. LAWS ch. 266, § 37E(a) (2015). Data security breach notification statutes require that companies and government entities encourage individuals to monitor their accounts for tampering if sensitive data is not kept confidential. N.C. GEN. STAT. § 75-65 (2015); CAL. CIV. CODE § 1798.82 (West 2015). Gramm-Leach-Bliley mandates notice requirements to allow bank customers to opt-out of permitted sharing of some of their financial information. Gramm-Leach-Bliley Act (GLBA), 15 U.S.C. § 6802 (2012).

76. Publication of location information can place persons, particularly police officers, cooperating defendants, and victims of stalking, in harm’s way. Grayson Barber, *Personal Information in Government Records: Protecting the Public Interest in Privacy*, 25 ST. LOUIS U. PUB. L. REV. 63, 77 (2006); Morrison, *supra* note 16, at 971. Some federal and state statutes offer privacy protection under limited circumstances to particular groups. *See, e.g.*, Family Educational Rights and Privacy Act (FERPA), 20 U.S.C. § 1232g(a)(5)(B) (2012) (mandating the option for parents to opt-out from the publishing of student directory information); CAL. GOV. CODE § 6254.21 (West 2015) (prohibiting the posting of any elected or appointed official’s home address or telephone number to the Internet without written permission).

77. Although Federal EEO law does not require non-disclosure of protected class status, Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e-2, some employers have developed best practices for avoiding related questions in order to provide a defense that they could not have based hiring decisions on information they did not have. In effect, privacy works as a barrier to discrimination. *See* Chad Derum & Karen Engle, *The Rise of the Personal Animosity Presumption in Title VII and the Return to “No Cause” Employment*, 81 TEX. L. REV. 1177 (2003) (arguing that a shift in the presumption of failure to hire for discriminatory reasons has occurred in favor of employers); Daniel J. Bugbee, *Employer’s Beware: Violating USERRA through Improper Pre-Employment*

and protection of intellectual development space for certain creative pursuits and for children are some of the interests protected under the umbrella of privacy law.<sup>79</sup> Other concepts associated with privacy include autonomy, dignity, and liberty.<sup>80</sup> Some of these privacy interests relate to rights against the government while others address privacy in the context of private relationships. In addition, although privacy is generally considered a personal interest, it is also advanced as an important benefit to society.<sup>81</sup>

Another point of debate is the authority for defining privacy interests. Some approaches embrace the idea that privacy is a personal choice.<sup>82</sup>

---

*Inquiries*, 12 CHAP. L. REV. 279 (2008) (discussing pre-employment inquiries under the Uniformed Services Employment and Reemployment Rights Act which protects those who served in the military).

78. Evidentiary privileges such as the attorney-client privilege are designed to encourage disclosure by providing confidentiality. MODEL CODE OF PROF'L CONDUCT R. 1.6 (1983).

79. The Children's Online Privacy Protection Act (COPPA) provides protections for children in the online environment, including parental consent requirements before certain personal information can be collected from a child. Children's Online Privacy Protection Act of 1998, Pub. L. No. 105-277, 112 Stat. 2681-728 (codified at 15 U.S.C. §§ 6501-6506 (2012)). Intellectual privacy is protected in state library privacy statutes and in the case of California, an e-reader privacy statute. California Online Privacy Protection Act (CalOPPA), CAL. BUS. & PROF. CODE § 22575 (West 2015).

80. Protections of some areas of personal integrity are recognized as constitutional freedoms from government intrusion, and this body or bodies of law are characterized as decisional privacy, information privacy, and/or liberty protections. *See, e.g.*, *Griswold v. Conn.*, 381 U.S. 479 (1965) (holding that the right of marital privacy was violated by a statute restricting the use of or provision of advice in support of contraception); *Whalen v. Roe*, 429 U.S. 589, 598-599 (1977) (noting privacy jurisprudence recognizes at least two types of interests, avoiding disclosure of personal information and independence in making certain kinds of important decisions); *Lawrence v. Texas*, 539 U.S. 538 (2003) (finding a Texas statute that criminalized sodomy intruded into the personal and private lives of individuals and violated the right to liberty under the Fourteenth Amendment). The Federal Trade Commission has committed to enforcing privacy promises even when the harm is not economic or physical or an unwanted intrusion, but merely unexpected disclosure of sensitive information about health or precise geolocation as well as less sensitive information such as purchase or employment history. FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS, 13-14 (March, 2012), <https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf>.

81. *See* Daniel J. Solove, "Nothing to Hide" and Other Misunderstandings of Privacy, 44 SAN DIEGO L. REV. 745 (2007).

82. Notice and choice were two of the first Fair Information Principles developed by a government advisory committee addressing privacy concerns in the United States. *See* Robert Gellman, *Fair Information Practices: A Basic History* (Feb. 11, 2015), <http://bobgellman.com/rg-docs/rg-FIPShistory.pdf> (documenting the history of the Fair Information Principles as well as their influence on U.S. laws).

Other approaches suggest privacy standards should reflect cultural norms,<sup>83</sup> and yet another view is that privacy might need to be imposed upon individuals by a paternalistic government.<sup>84</sup>

Disputes over the role of confidentiality also contribute to the instability of any comprehensive taxonomy. While privacy is generally considered to be about an individual's ability to avoid disclosure of his or her personal information, confidentiality is often used to describe a state of limited disclosure of that same information, perhaps to a person who has a duty to prevent further disclosure of conversations, such as an attorney providing legal advice. Some confidential relationships are recognized broadly throughout the law, while others are based on contractual principles or specific statutes that limit sharing of information.<sup>85</sup> Still other confidential relationships are supported by cultural, religious, or other social norms and have no enforcement mechanisms in the law.

---

83. The two-prong *Katz* test for violations of the Fourth Amendment includes both a subjective test for the defendant's expectation of privacy and an objective measure of the reasonable expectation of privacy. *Katz v. United States*, 389 U.S. at 347, 360–62 (Harlan, J., concurring). The reasonable expectation prong might well be an assessment of existing societal norms and realities. See Orin S. Kerr, *Four Models of Fourth Amendment Protection*, 60 STAN. L. REV. 503 (2007) (noting that societal understandings of privacy could be relevant to determining what constitutes Fourth Amendment reasonable expectations). In the online context, industry self-regulation for privacy is largely a measure of how much intrusion the market will tolerate without calling on Congress to formally regulate. See Omer Tene & J. Trevor Hughes, *The Promise and Shortcomings of Privacy Multistakeholder Policymaking: A Case Study*, 66 ME. L. REV. 437 (2014) (noting criticisms of industry codes of conduct and recommending structural supports to improve upon failed self-regulation).

84. See Anita L. Allen, *Coercing Privacy*, 40 WM. & MARY L. REV. 723, 755 (1999) (“Government will have to intervene in private lives for the sake of privacy and values associated with it.”).

85. Evidentiary privileges against compelled disclosures support several confidential relationships including attorney-client, spousal, clergy-penitent, and physician-patient relationships. Edward J. Imwinkelried, *THE NEW WIGMORE: A TREATISE ON EVIDENCE: EVIDENTIARY PRIVILEGES* §3.2.4 (2014); see also Neil M. Richards & Daniel J. Solove, *Privacy's Other Path: Recovering the Law of Confidentiality*, 96 GEO. L.J. 123 (2007) (promoting the common law claim of breach of duty of confidentiality). Consumer enforceability of privacy policies has not been successful because of an unclear contractual status and difficulty in proving harm. A variety of scholarly proposals have emerged. See, e.g., Woodrow Hartzog, *Reviving Implied Confidentiality*, 89 IND. L.J. 763 (2014); Joshua A. R. Fairfield, *“Do-Not-Track” as Contract*, 14 VAND. J. ENT. & TECH. L. 545 (2012). Much of privacy protection is conducted by the Federal Trade Commission through its authority to investigate and bring actions to address “unfair or deceptive trade practices,” which have yielded some penalties for violations of privacy promises. See Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

The role of practical obscurity adds another element to the conceptual framing of privacy. Like confidentiality, practical obscurity creates expectations of limited disclosure based on practical barriers to sharing rather than on legal or social restrictions. Most of the debate about confidentiality and practical obscurity relates to the role that the law should play in supporting these social or practical norms.<sup>86</sup>

Context plays an important role in defining privacy, confidentiality, and practical obscurity, yet this is a difficult factor to encapsulate in a taxonomy of sensitive information. The contextual approach emphasizes that privacy risks vary based on the circumstances in which information is shared, including the relationships between the sharer and recipient as well as their expectations at the time of sharing. Time is also a contextual factor that can have an impact on both the harms and benefits that attach to the disclosure of sensitive information. Some approaches to privacy embrace the idea that the value of privacy increases over time compared with other interests,<sup>87</sup> and yet in other instances privacy interests are treated as decreasing with the passage of time.<sup>88</sup>

## 2. *Charting the Piecemeal U.S. Approach to Privacy*

Another challenge in creating a taxonomy of sensitive information is that different information types are treated as sensitive in different areas of the law. The piecemeal approach evident in U.S. privacy law is a function of the federal system, a history of legislating in response to startling events,<sup>89</sup> and the balancing of interests promoted by stakeholders. Many

---

86. See Nissenbaum, *supra* note 14, 155–56 (describing privacy norms as a function of many variables and suggesting that “protecting privacy will be a messy task”).

87. The Court of Justice of the European Union held that Google must remove from its search results a link to a news article about a foreclosure that occurred more than a decade ago because the information was no longer timely. Case C-131/12, *Google Spain SL v. Agencia Española de Protección de Datos* (May 13, 2014), <http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&doclang=EN>.

88. See Douglas A. Kysar, *Kids & Cul-De-Sacs: Census 2000 and the Reproduction of Consumer Culture*, 87 CORNELL L. REV. 853, 870–75 (2001) (discussing the privacy concerns with census data collection despite assurances of confidentiality); HIPAA Privacy Rule, 45 C.F.R. § 164.502(f) (2015) (requiring covered entities to comply with the requirements of the HIPAA Privacy Rule for a period of fifty years following a decedent’s death).

89. For example, Congress passed the Video Privacy Protection Act (VPPA), 18 U.S.C. § 2710 (2012), in 1998 following the disclosure of Robert Bork’s video rental history during his nomination to the Supreme Court. See Neil M. Richards, *The Perils of Social Reading*, 101 GEO. L.J. 689, 694 (2013) (describing how “a horrified Congress quickly passed the VPPA,” perhaps upon realizing that politicians’ video rental records might otherwise be revealed as easily as Bork’s); Andrea Peterson, *How a Failed Supreme Court Nomination Is Still Causing Headaches for Hulu and Netflix*, WASH. POST (Dec. 27,

U.S. privacy laws protect particular types of information within the context of a regulated sector, such as health care and banking, or within the context of limiting government power. The result is that particular information types may be protected in one sector although their privacy benefits can be outweighed by competing interests in another. Activity at the state level has also resulted in multiple approaches in areas such as data security breach notification requirements, and so the fragmentation is ongoing and pervades the U.S. legal system.

A related issue for the creation of a comprehensive taxonomy is that some sensitive information types are more clearly defined by law than others. Some laws are grounded in general principles like “unfair or deceptive trade practices”<sup>90</sup> or tautologies found in common law torts that provide civil remedies for the disclosure of “private facts.”<sup>91</sup> These vague constitutional and tort protections for privacy contrast with health regulations such as the Health Insurance Portability and Accountability Act (HIPAA) that contains a list of seventeen sensitive information types that need to be redacted before regulated health entities can share personal health information.<sup>92</sup> Some privacy-related laws do not define sensitive information types at all and instead draw on influential policy statements,<sup>93</sup> industry standards,<sup>94</sup> and other areas of law.<sup>95</sup> Some U.S.

---

2013), <http://www.washingtonpost.com/blogs/the-switch/wp/2013/12/27/how-a-failed-supreme-court-bid-is-still-causing-headaches-for-hulu-and-netflix>.

90. Federal Trade Commission (FTC) Act, 15 U.S.C. § 45 (2012). Much of privacy law now comes from the FTC’s investigative and enforcement powers to bring or settle lawsuits when companies under their jurisdiction arguably fail to live up to their privacy promises. See Solove & Hartzog, *supra* note 85.

91. RESTATEMENT (SECOND) OF TORTS § 652D (1977) (“One who gives publicity to a matter concerning the private life of another is subject to liability to the other for invasion of his privacy, if the matter publicized is of a kind that (a) would be highly offensive to a reasonable person, and (b) is not of legitimate concern to the public.”).

92. HIPAA Privacy Rule, 45 C.F.R. § 164.514(b)(2) (2015). The rule includes an eighteenth requirement, to remove “[a]ny other unique identifying number, characteristic or code.” 45 C.F.R. § 164.514(b)(2)(i)(R) (2015).

93. See Gellman, *supra* note 82.

94. See *PCI SSC Data Security Standards Overview*, PCI SECURITY STANDARDS COUNSEL, [https://www.pcisecuritystandards.org/security\\_standards/index.php](https://www.pcisecuritystandards.org/security_standards/index.php) (last visited Apr. 3, 2015).

95. See *Nat’l Archives & Records Admin. v. Favish*, 541 U.S. 157, 167–68 (2004) (finding that the Freedom of Information Act exemption for personal privacy extends to the familial concern in controlling the deceased’s death images, in accordance with common law notions of privacy). Some state court electronic filing rules refer to state identity theft and breach notice laws, such as in North Carolina. See *N.C. eFiling Rules*, *supra* note 10. Rule 6.3 for e-filing in North Carolina defines “private information,” as including sensitive, personal, or identifying information which must be excluded or

privacy laws are hybrids, with illustrative but non-exhaustive lists of protected information types.<sup>96</sup> Sometimes information is protected from some uses but not others, as in the Fair Credit Reporting Act, which limits the circumstances under which a consumer reporting agency can distribute consumer credit reports.<sup>97</sup>

Whatever benefits the sectoral approach brings, they are increasingly threatened by the ease with which information can be shared and aggregated. The increase in data brokers and the work of computer scientists and journalists highlight the leaky boundaries between separately regulated sectors and the potential for recreating previously redacted information by merging separate databases.<sup>98</sup> Information not restricted from disclosure in one context can obviate privacy protections in other parts of the dynamic information ecosystem. This development affects not just those individuals whose sensitive information is exposed through one sector but also those industry actors who invest in costly privacy and security approaches that prove to be ineffective. Public records in particular can spoil the privacy protections required in other areas because

---

partially described in court documents. The statutory basis for Rule 6.3 is N.C. Gen. Stat. § 132-1.10(d).

96. See Driver's Privacy Protection Act (DPPA), 18 U.S.C. §§ 2721–2725 (2012); *Dahlstrom v. Sun-Times Media*, 777 F.3d 937 (7th Cir. 2015) (holding that the DPPA's prohibition on disclosure of personal information in driving records did not raise heightened First Amendment scrutiny). The Freedom of Information Act lists exemptions, but refers to "personal privacy" somewhat unhelpfully. 5 U.S.C. § 552 (2012).

97. Fair Credit Reporting Act, 15 U.S.C. § 1681b (2012).

98. Journalists have been able to identify "anonymous" Internet users through records of their search history. Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), <http://query.nytimes.com/gst/fullpage.html?res=9E0CE3DD1F3FF93AA3575BC0A9609C8B63>. Researchers have used "anonymous" Netflix viewing information released by the company to re-identify some of its customers. Steve Lohr, *Netflix Cancels Contest Plans and Settles Suit*, N.Y. TIMES (Mar. 12, 2010), <http://bits.blogs.nytimes.com/2010/03/12/netflix-cancels-contest-plans-and-settles-suit>. A graduate computer science student at MIT was able to re-identify the Governor of Massachusetts William Weld through presumptively anonymized state hospital records; the student recently reported forty percent re-identification capabilities in most contexts. Latanya Sweeney et al., *Identifying Participants in the Personal Genome Project by Name* (Apr. 29, 2013), [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2257732](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2257732). The FTC's 2014 report on Data Brokers outlined a growing industry of data collectors and resellers who intermingle public records, information on the web, and proprietary data. FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY (2014), <https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>.

they provide information that can be used for re-identifying individuals or for connecting a profile with information that carries privacy risks.

With these concerns in mind, we set out to create a sensitive information taxonomy that would allow for the identification of sensitive information in court records. We utilized a taxonomic approach for several reasons. First, we wanted to study the frequency of sensitive information in court records without first making a normative claim about what information types should be considered private. As a result, we cast the net widely and included a large number of information types, even those that appeared to have only a modicum of support in existing privacy law and scholarship. Second, we believe that the process we undertook to create our taxonomy will be valuable to other scholars and policymakers. In the sections that follow, we describe how we created our taxonomy and why we chose the information types that we did. Of course, not everyone will agree with our final list. Nevertheless, our taxonomy has proven to be a helpful guide in the assessment of the privacy risks that can arise from the public disclosure of court records, especially at a time when privacy laws have limited or unclear application to such records. Finally, even for those who disagree with our inclusions and exclusions, the instant taxonomy will serve as a useful starting point for the development of alternative taxonomies that scholars can apply to other information contexts.

#### B. CRITERIA FOR INCLUSION IN THIS STUDY

This project's taxonomy of sensitive information represents a broad list of information types that are protected by U.S. privacy law or that have been identified by scholars or others as information that should be protected from public disclosure.<sup>99</sup> To facilitate coding and analysis of the court records in our study, we grouped the various sensitive information types into the following thirteen categories:

1. Assets
2. Civil Proceedings
3. Computer Use
4. Criminal Proceedings
5. Education
6. Employment
7. Financial Information
8. Health

---

99. We conducted a survey of federal and state constitutional, tort, statutory, and regulatory law as well as federal and state court rules, European law, and legal scholarship.

9. Identity
10. Images
11. Intellectual Pursuits
12. Location
13. Sexual Activities

These are thematic categories that capture similarities in subject matter for the 140 sensitive information types that we searched for in the court records. Some information types logically fit in multiple categories, but we placed them in one category as a matter of simplicity for the coding and analysis of the records. There are, of course, other ways the individual information types can be categorized. We describe the makeup of each category in the sections that follow and provide a full listing of all of the coded information types in the Appendix.

After initial testing of the taxonomy, we decided to limit recording of sensitive information types to those occurrences that the coder could associate with an identified individual. In other words, we only coded for sensitive information that was associated with a person named in full or by last name within the brief or appendices of each document in the study. The identified individual did not have to be named on the same page where the sensitive information type occurred, but the association had to be clear to the coder from the information within the document.<sup>100</sup> For example, we would code for “Anne Klinefelter’s Browning semi-automatic handgun” because it is apparent from the document that the gun is owned or possessed by Anne, but we would not code for “a Browning semi-automatic handgun was found in the street outside the grocery store” because the information is not associated with an identified individual.<sup>101</sup>

---

100. The only exception we made was for social security numbers because of their utility as a stand-in for personal identification. *See infra* note 139 and accompanying text.

101. Our requirement that sensitive information types needed to be associated with individuals named within the court briefs means that our findings do not include data that might directly support considerations of how discrete appearances of sensitive information (not associated with persons identified in court records) might be linked to individuals when court records are read in conjunction with outside sources. *See* FED. TRADE COMM’N, PROTECTING CONSUMER PRIVACY, *supra* note 80, at 35–38 (suggesting that non-personally identifying information (PII) can be increasingly transformed into personally identifying information through re-identification); Christopher Wolf, *Technological Advances and Privacy Challenges*, in UNDERSTANDING DEVELOPMENTS IN CYBERSPACE LAW 23 (2014) (advising that traditional solutions focused on personally identifying information are undercut by big data practices and should be supplemented with techniques such as measuring risk of re-identification, and noting that publicly released data present greater risks for re-identification than data not publicly released); Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a*

We did not record occurrences of names alone, other than the names of minor children, absent some connection between the name and another piece of sensitive information (e.g., a person named as a juror, witness, or rape victim).<sup>102</sup> Where an individual was associated with another piece of sensitive information, we coded for whether the individual was an “adult,” “minor,” or “unknown.” In privacy law, minors receive more protection than adults, so this status is itself a piece of sensitive information.<sup>103</sup> We did not code for information types associated with entities such as businesses, associations, and other groups.

What follows is an overview of the information types and categories in our taxonomy and some description of the sources that we used to create the taxonomy.

### 1. *Assets*

The “assets” category contains information relating to an identified person’s possession or ownership of assets that might be considered sensitive, including financial assets, real estate, and vehicle identification numbers and license plate numbers. It also contains information indicating that an individual has a gun permit, filed a gun permit application, or possessed or owned a gun.

Scholars have noted that property ownership, including real estate ownership or rental status, constitutes information that can be used to identify individuals and that enables the creation of profiles used by data aggregators.<sup>104</sup> The Federal Trade Commission has noted that property records maintained by states are part of a growing data broker industry

---

*New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1836–47 (2011) (noting that non-PII are no longer immutable categories due to the risk of re-identification and proposing an assessment using a continuum of identification risk).

102. We tested the coding of every appearance of a name early in the project, but so many names appeared in the briefs and appendices that it threatened to overwhelm our resources. We decided to leave this particular type of examination of court records to other researchers.

103. See, e.g., Children’s Online Privacy Protection Act (COPPA), 15 U.S.C. §§ 6501–06 (2012); REPORTERS COMM. FOR FREEDOM OF THE PRESS, PRIVATE EYES: CONFIDENTIALITY ISSUES AND ACCESS TO POLICE INVESTIGATION RECORDS 4–5 (2010), <http://www.rcfp.org/rcfp/orders/docs/PRIVATEEYES.pdf> (discussing protections for juvenile records); CAL. BUS. & PROF. CODE § 22581 (2015) (protecting minors online). Court electronic filing rules also generally allow for the replacement of the names of minors with their initials. See *supra* note 10.

104. Helen Nissenbaum, *Privacy in an Information Age: The Problem of Privacy in Public*, 17 LAW AND PHILOSOPHY 559, 561, 577 (1998) (listing information types traditionally treated as public and increasingly being used in the digital environment as identifying information in the organized surveillance of individuals).

that creates some risks of harm for consumers.<sup>105</sup> Vehicle identification numbers and license plate numbers are used in similar ways and are restricted under some state laws as well as the federal Driver's Privacy Protection Act.<sup>106</sup>

Information on gun ownership and possession is included because some states have passed legislation to protect the privacy of gun owners. For example, North Carolina exempts gun registration records from its public records law which would otherwise require public access to that information.<sup>107</sup> The Florida Firearm Owners Privacy Act limits a physician's ability to inquire about firearm access in patient interviews.<sup>108</sup> In addition, scholars raise the possibility that public disclosure of gun registration unconstitutionally burdens the right to bear arms.<sup>109</sup>

## 2. *Civil Proceedings*

The "civil proceedings" category is an organizing point for a number of types of information that relate to civil lawsuits and other non-criminal judicial proceedings. This category includes information relating to adoption, child support, civil commitment to a penal or mental facility, custody or guardianship proceedings, information indicating that an

---

105. FED. TRADE COMM'N, DATA BROKERS *supra* note 98, at 11–12 (reporting how state and local government records including property records are collected by data brokers either directly or indirectly); *see also* FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY, *supra* note 80, at 69 (recommending legislation to improve transparency in the data broker industry).

106. For example, North Carolina limits the sharing of vehicle identification numbers acquired through toll road administration. N.C. GEN. STAT. § 136-89.213 (2015). Regulations implementing the HIPAA Privacy Rule prohibit the sharing of vehicle identifiers and serial numbers, including license plate numbers, connected to personal health information. 45 C.F.R. § 164.514(b)(2)(i)(L) (2015). The federal Driver's Privacy Protection Act, 18 U.S.C. §§ 2721–2725 (2012), limits state departments of motor vehicles from sharing personal information except for specified purposes. The statute has been read to curtail distribution of "personal information," which is defined broadly. *See Dahlstrom v. Sun-Times Media*, 777 F.3d 937, 943 (7th Cir. 2015) ("[T]he DPPA's language appears broad: personal information means information that identifies an individual, . . . and there is no indication that Congress intended the enumerated list of examples to be exhaustive." (internal quotations omitted)).

107. N.C. GEN. STAT. §§ 14-415.17(c), 14-405(b), 14-406(a) (2015).

108. *See Wollschlaeger v. Governor of Florida*, 760 F.3d 1195 (11th Cir. 2014) (finding constitutional the effect of the Florida Firearm Owners Privacy Act on physicians' free speech rights); Act of April 26, 2011, 2011 Fla. Laws 112 (codified at FLA. STAT. §§ 381.026, 456.072, 790.338).

109. *See, e.g., Eugene Volokh, Implementing the Right to Keep and Bear Arms for Self-Defense: An Analytical Framework and a Research Agenda*, 56 UCLA L. REV. 1443, 1548–49 (2009).

identified individual was the subject of a dependency or neglect proceeding, party to a divorce, juror name, and prior adverse civil judgments.

Information that falls within this category may be regarded as sensitive because individuals have no choice but to share these types of personal information in order to make use of government services or to remain law-abiding citizens.<sup>110</sup> The particular privacy implications of each of these information types are also advanced because of the special dignitary harms or, in the case of juror names, risk of retaliatory harms from public disclosure.<sup>111</sup> Information arising in civil proceedings that falls into a more specific category such as financial or health information is included in those more detailed categories below.

### 3. *Computer Use*

A number of information types that relate to an individual's use of computers or electronic information services comprise the "computer use" category: Instant Messenger or SMS identifier; IP address; Internet search history; Internet Service Provider (ISP) records including account number, billing information, or online access logs; computer password; Radio Frequency Identification (RFID); screen or user name for accessing a website or other online service; and Voice Over Internet (VOIP) username or number. These information types are culled from several different federal and state statutes as well as scholarship advocating protection for this kind of information.<sup>112</sup>

---

110. See Grayson Barber, *Personal Information in Government Records: Protecting the Public Interest in Privacy*, 25 ST. LOUIS U. PUB. L. REV 63, 71–72 (2006) (suggesting that court records "often contain information that is exquisitely personal"); *Whalen v. Roe*, 429 U.S. 589, 605–06 (1977) (holding that a state database on individuals with prescriptions for controlled substances did not violate a right to privacy because the data was not for public disclosure and was kept secure); Grayson Barber & Frank L. Corrado, *Public Access to Government Records and How Transparency Protects Privacy*, N.J. LAW., Oct. 2011, at 60 (protesting the government practice of selling personal information, and advocating transparency in order to generate citizen advocacy for greater privacy protection).

111. See Kenneth J. Melilli, *Disclosure of Juror Identities to the Press: Who Will Speak for the Jurors?*, 8 CARDOZO PUB. L. POL'Y & ETHICS J. 1 (2009) (advocating confidentiality of jury service to protect former jurors from harassment and physical threats); Kristin A. Henderson, *Lessons from Bankruptcy Court Public Records*, 23 LEGAL REFERENCE SERVS. Q. 55, 73, 76–77 (2004) (evaluating the particular sensitivity of information provided in bankruptcy proceedings).

112. See, e.g., Children's Online Privacy Protection Act (COPPA), C.F.R. §§ 312.1–2; Omer Tene & Jules Polonetsky, *To Track or "Do Not Track": Advancing Transparency and Individual Control in Online Behavioral Advertising*, 13 MINN. J.L. SCI. & TECH.

#### 4. *Criminal Proceedings*

Like the civil proceedings category, the “criminal proceedings” category serves as an organizing device for sensitive information related to the justice system. For this category, the information types are associated with law enforcement and criminal judicial proceedings, including information that identifies an individual as the subject of a criminal investigation, arrest, incarceration, conviction, sentence, or parole. The category also includes mug shots and pre-sentence investigation reports, sexual abuse allegations, child abuse allegations, and information concerning charges or convictions arising in juvenile proceedings. Additional information types are included for juror name, domestic violence victim name, rape victim name, and other crime victim name. The criminal proceedings category also includes cooperating defendant name, informant name, and witness name.

The information types we have listed within this category are widely regarded as sensitive.<sup>113</sup> For example, many scholars assert that the public disclosure of the names of crime victims and witnesses leads to the further victimization of those who have suffered from or witnessed criminal activity.<sup>114</sup> Others point to the stigma that attaches to individuals who have been subjected to criminal investigation, charge, or conviction.<sup>115</sup>

---

281, (2012); Chris J. Hoofnagle et al., *Behavioral Advertising: The Offer You Cannot Refuse*, 6 HARV. L. & POL'Y REV. 273 (2012).

113. See, e.g., JAMES B. JACOBS, THE ETERNAL CRIMINAL RECORD 54–69 (2015) (describing the many types of criminal information in court records and criticizing their widespread availability); Sadiq Reza, *Privacy and the Criminal Arrestee or Suspect: In Search of a Right, In Need of a Rule*, 64 MD. L. REV. 755 (2005) (proposing increased privacy protections for the criminally-accused); Deanna K. Shullman & Mark R. Caramanica, *Mug Shots on Lockdown: Government and Citizen Backlash to “Exploitation” Websites Surges, Free Speech is the Casualty*, 30 COMM. LAW. 13 (2014) (surveying responses to businesses offering to take down mug shots for a fee and examining a split in federal circuit courts on the constitutionality of restrictions); Rebecca Hulse, *Privacy and Domestic Violence in Court*, 16 WM. & MARY J. WOMEN & L. 237 (2010) (examining the privacy rights of domestic violence victims in court and concluding that special protections should extend beyond family court contexts); Morrison, *supra* note 16 at 921 (highlighting the risks of harm from online court records in criminal cases and recommending the redaction of names of cooperating defendants and other informants while increasing transparency in the use of these law enforcement practices).

114. See, e.g., Joel M. Schumm, *No Names, Please: The Virtual Victimization of Children, Crime Victims, the Mentally Ill, and Others in Appellate Court Opinions*, 42 GA. L. REV. 471, 486–93 (2008).

115. See, e.g., Devah Pager, Bruce Western & Naomi Sugie, *Sequencing Disadvantage: Barriers to Employment Facing Young Black and White Men with Criminal Records*, 623 ANNALS AMER. ACAD. POL. & SOC. SCI. 195, 199 (2009), <http://ann.sagepub.com/content/623/1/195> (finding that men with a felony drug conviction were fifty percent less

The combination of online and data broker exposure of often stale and incomplete arrest and conviction information is criticized as creating long-term barriers to fresh starts including negative impacts on “employment and housing prospects, parental rights, educational opportunities, freedom of movement, and just about every other aspect of daily life.”<sup>116</sup> Broad exposure of and reliance on records of criminal activity is said to permanently mark ex-offenders as outlaws and restrict their ability to forge a path outside of crime, “a terrible outcome for society.”<sup>117</sup> Information arising in juvenile delinquency proceedings has long been considered to be particularly sensitive because the disclosure of such information was “thought to hinder their rehabilitation by impairing their relations with the community [and] by stigmatizing them such that they view themselves as wrongdoers and act accordingly.”<sup>118</sup>

As a result, a number of states have adopted or are considering broad sealing and expungement laws for various types of criminal information.<sup>119</sup> Although some of these laws will likely face significant constitutional challenges,<sup>120</sup> there is clearly a concerted effort by privacy and criminal justice advocates to limit the public disclosure of many types of criminal information.

##### 5. *Education*

The “education” category encompasses five information types that relate to students at all levels of the education system: income eligibility for the National School Lunch program, the amount of financial aid awarded from federal or private sources, information indicating that a student was disciplined by his or her school, grades or other feedback from a school about a student’s performance, and student identifiers.

Educational information is generally regarded as sensitive because it relates to a vulnerable class of individuals, often minors, who must share information with educational institutions, sometimes in a compulsory

---

likely than men without any record to receive a callback or be offered an entry-level job; black men with a record who applied were twice as likely as white men to be saddled with this “criminal record penalty”).

116. See, e.g., Jenny Roberts, *Expunging America’s Rap Sheet in the Information Age*, 2 WIS. L. REV. 321, 327 (2015).

117. JACOBS, *supra* note 113, at 306.

118. Reza, *supra* note 113, at 785.

119. See Roberts, *supra* note 116, at 322.

120. See *supra* notes 25–30 and accompanying text.

education context. The information types included in this category are drawn from several federal statutory protections.<sup>121</sup>

#### 6. *Employment*

Three information types are included in the “employment” category: information that an individual was disciplined by an employer, information describing an individual’s military discharge, and performance evaluations of an employee. Information about the location where an individual works is included in the “location” category.

Employee privacy requirements vary by jurisdiction under statutory and common law.<sup>122</sup> In some states, performance evaluations are exempt from public disclosure.<sup>123</sup> The federal Freedom of Information Act includes an exemption for disclosure of contents of personnel files if that information would constitute a clearly unwarranted invasion of personal privacy,<sup>124</sup> and this exemption has been applied to performance appraisals.<sup>125</sup> Private employee privacy in performance appraisals varies, but in some states this information is protected by statute or by common law.<sup>126</sup>

121. The Family Educational Rights and Privacy Act protects student education records of institutions receiving federal funding. Family Educational Rights and Privacy Act, 20 U.S.C. § 1232(g) (2012); 34 C.F.R. § 99 (2015). The National School Lunch Act protects the confidentiality of the names of individual students who qualify for school lunch assistance. 42 U.S.C. § 1758(6) (2012).

122. See Pauline T. Kim, *Privacy Rights, Public Policy, and the Employment Relationship*, 57 OHIO ST. L.J. 671 (1996); *Access to Social Media Usernames and Passwords*, NAT’L CONFERENCE OF STATE LEGISLATURES, <http://www.ncsl.org/research/telecommunications-and-information-technology/employer-access-to-social-media-passwords-2013.aspx> (last visited Apr. 3, 2015) (listing recent legislation intended to strengthen workplace privacy regarding personal employee social media and other accounts); N.C. GEN. STAT. §§ 132-1.2, 160A-168 (2015).

123. States take different approaches to the accessibility of public employees’ performance evaluations. See Roger A. Nowadsky, *A Comparative Analysis of Public Records Statutes*, 28 URB. LAW. 65, 86 (1996) (surveying states’ laws and finding that in most states personnel files are presumptively private).

124. 5 U.S.C. § 552(b)(6) (2012).

125. *McLeod v. U.S. Coast Guard*, No. 96-5071, 1997 WL 150096 (D.C. Cir. 1997) (finding privacy interest in Coast Guard officer’s evaluation report); *Smith v. Dep’t of Labor*, 798 F. Supp. 2d 274, 284–85 (2011) (holding that disclosure of records containing performance appraisal information would constitute an unwarranted invasion of employee’s personal privacy.)

126. See Laura B. Pincus and Clayton Trotter, *The Disparity Between Public and Private Sector Employee Privacy Protections: A Call For Legitimate Privacy Rights for Private Sector Workers*, 33 AM. BUS. L.J. 51, 54 (1995) (comparing states’ approaches under statute and common law).

### 7. *Financial Information*

The “financial information” category contains a number of information types that relate to a person’s financial condition and accounts. Separate types were specified for the account numbers associated with an individual’s savings, checking, or other financial account; loan account numbers; credit card numbers; debit card numbers; and other types of financial accounts not already specified. Other information types in this category include information indicating that a person filed for bankruptcy or was adjudged to be bankrupt, that a person has been the subject of a foreclosure judgment, that a person owes a debt, or that a person has a lien on assets due to unpaid taxes. The ownership of physical financial assets such as stock certificates, cash, and coins is included in the “assets” category.

Tax returns are listed in this category as is information about compensation in the form of salary, wage, or other financial benefits, including stock options, court ordered payments, and other forms of compensation.<sup>127</sup> Insurance policy numbers, credit reports, and an individual’s status as an identity theft victim are also considered sensitive information types under various laws.

Information that falls within this category is regarded as sensitive because it not only reveals details about a person’s net worth, but it also may be useful in the commission of identity theft and consequential financial theft or credit harm. States have passed varying forms of restrictions on the sharing of financial and identifying information in order to reduce the risk of identity theft.<sup>128</sup> Other statutes such as the federal Fair Credit Reporting Act and the Gramm-Leach-Bliley Act provide protections for consumers seeking to restrict sharing of financial and related information.<sup>129</sup> In addition, debt and bankruptcy can result in

---

127. See Cynthia Blum, *The Flat Tax: A Panacea for Privacy Concerns?*, 54 AM. U. L. REV. 1241, 1262–81 (2005) (outlining a variety of harms that could result from inappropriate uses of tax information and recommending safeguards against disclosure and misuse).

128. The National Conference of State Legislatures maintains a chart of state data security breach notification laws intended to provide individuals an opportunity to monitor credit and financial accounts and to change passwords and credit card numbers to minimize the potential for identity-theft and related harms. *Security Breach Notification Laws*, NAT’L CONFERENCE OF STATE LEGISLATURES, <http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx> (last updated Jan. 4, 2016).

129. Fair Credit Reporting Act, Pub. L. 91-508, Title VI, § 601, 84 Stat. 1114, 1128 (1970) (codified as amended at 15 U.S.C. §§ 1681–1681x (2012)); Gramm-Leach-Bliley

negative treatment even after an individual has regained financial stability.<sup>130</sup> The evolving law of court rules for electronic filing generally requires redaction of financial account numbers.<sup>131</sup>

### 8. *Health*

The “health” category includes information about abortion, cause of death, place of death, communicable diseases, dates of a hospital stay, disability status, drug or alcohol dependency, drug or alcohol treatment, HIV/AIDS status, and information relating to prescription medications. Paternity test information and pregnancy information are both in this category. Also included in this category are health plan beneficiary numbers, medical billing numbers, medical device identifiers or serial numbers, and medical record numbers. Other information in this category includes health diagnosis or treatment information not previously specified, genetic information, and medical conditions that are not the subject of diagnosis or treatment by a health care professional.

Information that falls within this category is regarded as highly sensitive because of the potential for discrimination based on perceptions of reduced capabilities or assumptions about unpopular causal behaviors.<sup>132</sup> Confidentiality of health information shared with a physician dates at least as far back as the Hippocratic Oath,<sup>133</sup> and is protected by many privilege laws and tort liability in some situations.<sup>134</sup>

The sources for information under the health category include the regulations authorized by HIPAA and its amendments,<sup>135</sup> privacy tort

---

Act, Pub. L. No. 106-102, § 527(4), 113 Stat. 1338, 1449 (1999) (codified as amended at 15 U.S.C. § 6827(4) (2012)).

130. Negative treatment of an individual based on debt that has been settled is sometimes considered an inappropriate response that should be prevented through restrictions in access to information about the debt. The right to be forgotten decision in Spain addressed this issue. See *supra* note 87.

131. See *supra* note 10.

132. See Charity Scott, *Is Too Much Privacy Bad for Your Health?* 17 GA. ST. U. L. REV. 481, 491–95 (2000) (reporting overwhelming public support for health privacy and articulating the harms of privacy violations as well as some benefits for certain exceptions).

133. Oath and Law of Hippocrates, circa 400 B.C.

134. Most states recognize an evidentiary privilege for physician-patient communications. See Edward J. Imwinkelried, *THE NEW WIGMORE: EVIDENTIARY PRIVILEGES* § 6.2.6 (2014); *McCormick v. England*, 494 S.E.2d 431, 437 (S.C. Ct. App. 1997) (reviewing the law of other states and joining the majority to recognize a tort for a “physician’s breach of the duty to maintain the confidences of his or her patient in the absence of a compelling public interest or other justification for the disclosure”).

135. The HIPAA Privacy Rule prevents the sharing of personal health information unless it is de-identified through the redaction of seventeen identifiers or through some

cases,<sup>136</sup> and federal constitutional law suggesting that information privacy may apply to some prescriptions.<sup>137</sup> The federal Genetic Information Nondiscrimination Act provides regulatory protection for genetic information.<sup>138</sup>

### 9. *Identity*

The “identity” category contains a number of information types that relate to an individual’s physical and other characteristics that allow others to identify the individual. Like the other information types that we coded, information in this category must be associated with an identified individual. We did not code for the occurrence of full names or last names alone, even though we did use the appearance of associated names as the qualifying factor for almost all of the information types. We coded for Social Security Number (SSN) with or without names because SSNs are unique identifiers on their own.<sup>139</sup> We found that tracking names was simply too coder-intensive and provided inconsistent levels of information given that some names are common in the population and others are not. We did, however, code for the name of a minor child if it appeared in the

---

approved statistical approach. HIPAA Privacy Rule, 45 C.F.R. § 164.514(b)(2) (2015). The Driver’s Privacy Protection Act prohibits states from selling or otherwise sharing drivers’ “highly restricted personal information.” 18 U.S.C. § 2721 (2012).

136. See, e.g., *Byrne v. Avery Ctr. for Obstetrics and Gynecology*, 314 Conn. 433 (2014); Robert H. Thornburg, *Florida Privacy Law: Potential Application of Intentional Tort Principles and Florida’s Constitutional Right of Privacy as Safeguards to Governmental and Private Dissemination of Private Information*, 4 FLA. COASTAL L.J. 137 (2003).

137. See *Whalen v. Roe*, 429 U.S. 589, 605 (1977) (holding that a New York database containing the names of individuals who were prescribed a controlled drug to treat depression did not burden a constitutional right to privacy because the statute had a rational basis and because the state adopted reasonable data security, but suggesting that a right to avoid wide disclosure of prescription information was at issue).

138. Genetic Information Nondiscrimination Act (GINA) of 2008, Pub. L. No. 110-233, 122 Stat. 881; 20 C.F.R. § 1635.1–12 (2015).

139. For a discussion of why we did not code for names alone, see *supra* note 102 and accompanying text. The perceived utility of social security numbers in facilitating identity theft and financial harm has inspired many states to pass legislation to restrict government collection of this information. See, e.g., N.C. GEN. STAT. § 132-1.10(d) (“No person preparing or filing a document to be recorded or filed in the official records of the register of deeds, the Department of the Secretary of State, or of the courts may include any person’s social security, employer taxpayer identification, drivers license, state identification, passport, checking account, savings account, credit card, or debit card number, or personal identification (PIN) code or passwords in that document, unless otherwise expressly required by law or court order, adopted by the State Registrar on records of vital events, or redacted.”).

records because privacy laws provide special protection for children in a variety of contexts.<sup>140</sup>

The identity category includes driver's license number, email address, fax number, mother's maiden name, passport number, city and state of birth, professional certificate or license number, state identification number, and telephone number. Although we initially sought to code for gender,<sup>141</sup> we ultimately dropped this information type because it became unworkable. English language and naming conventions make gender cues so numerous that it was overwhelming the coding process.<sup>142</sup> We kept an information type for gender identity change, even though that designation is itself a contested and complicated issue.

This category also includes a number of information types associated with biological traits. Age, date of birth, date of death, barefoot print, fingerprint, gait, iris print or recognition, and voice print are included here, as are racial or ethnic origin. Some of this biological information is considered sensitive under a number of laws and is the subject of scholarship advocating increased privacy protection relating to the collection and use of biometric information.<sup>143</sup> Dates of birth and death

---

140. The Children's Online Privacy Protection Act provides special requirements for the collection of identifying information from children under the age of thirteen including first and last name. 16 C.F.R. §§ 312.1–3 (implementing 15 U.S.C. §§ 6501–6508). Federal Model Rule of Appellate Procedure 5.2(a)(3) provides that minor children may be identified with initials only. The North Carolina e-filing rules permits minors to be "identified by initials," *N.C. eFiling Rules*, *supra* note 10, at Rule 6.3; *see also* Schumm, *supra* note 114.

141. Federal and state statutes prohibit discrimination on the basis of gender in contexts such as employment and housing. *See, e.g.*, Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e (prohibiting employment discrimination based on race, color, national origin, sex, and religion); Fair Housing Amendments Act of 1988, 42 U.S.C. § 3604 (prohibiting housing rental or sale on the basis of certain traits including sex). Constitutional protections have also been recognized for gender nondiscrimination. *See* *Craig v. Boren*, 429 U.S. 190 (1976) (applying intermediate scrutiny to gender discrimination). While access to gender information is generally not restricted, requests for information relating to gender can create vulnerability for discrimination claims, so this information has been considered sensitive.

142. Pronouns alone triggered huge numbers of coding opportunities and created confusion about their meaning as neutral or gender-aware applications. In addition, coders might misread some names as conveying gender information.

143. The National Institute of Standards and Technology (NIST), an agency of the U.S. Department of Commerce, states, "Biometric technologies are used to establish or verify personal identity against previously enrolled individuals based upon recognition of a physiological or behavioral characteristic. Examples of biological characteristics include hand, finger, facial, and iris. Behavioral characteristics are traits that are learned or acquired, such as dynamic signature verification and keystroke dynamics." *Biometric Standards Program and Resource Center*, NIST, <http://www.nist.gov/itl/csd/scm/>

are restricted from disclosure under regulations implementing HIPAA. State identity-theft protection acts and court rules for electronic filing systems tend to require redaction of full birth dates.<sup>144</sup> Racial and ethnic origin define groups that are protected under the Fourteenth Amendment and by nondiscrimination statutes that generally place limitations on access to or use of this information.<sup>145</sup>

### 10. Images

The “images” category captures occurrence of photographs and video that contain a full-frontal view of an individual’s face or features; show sexual organs of an undressed individual; show a person in a state of partial undress indicated to be taken without their consent; and photographs or videos depicting violence, abuse, or death of an individual. Video recordings that depict sexual acts are included in the “sexual activities” category.

These information types are drawn from protections provided through privacy torts, state statutes, and scholarship advocating additional protections for images, especially with the growing use of facial recognition.<sup>146</sup>

---

biometric-standards.cfm (last updated March 19, 2015); *see also* Margaret Hu, *Biometric ID Cybersurveillance*, 88 IND. L.J. 1475 (2013); Laura K. Donohue, *Technological Leap, Statutory Gap, and Constitutional Abyss: Remote Biometric Identification Comes of Age*, 97 MINN. L. REV. 407 (2012). Regulations implementing HIPAA restrict distribution of personal health information that contains biometric identifiers including finger and voice prints. 45 C.F.R. § 164.514(b)(2)(i)(P) (2015); *see also* N.C. GEN. STAT. § 14-113.20(b)(11) (2015).

144. *See* 45 C.F.R. § 164.514(b)(2)(i)(C); North Carolina Identity Theft Act, N.C. GEN. STAT. § 14-113.20.

145. *See* U.S. CONST. amend. XIV; *Adarand Constructors, Inc. v. Peña*, 515 U.S. 200, 201–02 (1995); *Gratz v. Bollinger*, 539 U.S. 244, 249–50 (2003); *see also* Title VII of the Civil Rights Act of 1964, 42 U.S.C. § 2000e (2012) (prohibiting employment discrimination on the basis of race, color, national origin, sex, and religion); Age Discrimination in Employment Act, 29 U.S.C. §§ 621–634 (2012) (prohibiting age discrimination against individuals over forty). States may also have nondiscrimination laws. While these statutes may not prohibit the gathering of information related to the protected traits, a strong defense against discrimination claims is that the sensitive information was not accessed.

146. *See, e.g.*, N.C. GEN. STAT. § 132-1.8. (protecting the confidentiality of photographs and video or audio recordings made pursuant to autopsy). Nonconsensual sharing of nude photographs or images of sexual activity are the subject of privacy torts claims and new statutes. *See* Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345 (2014); Clay Calvert & Justin Brown, *Video Voyeurism, Privacy, and the Internet: Exposing Peeping Toms in Cyberspace*, 18 CARDOZO ARTS & ENT. L.J. 469 (2000); Jeffrey R. Boles, *Documenting Death: Public Access to Government Death Records and Attendant Privacy Concerns*, 22 CORNELL J.L. &

### 11. *Intellectual Pursuits*

The “intellectual pursuits” category is a catch-all category that covers a range of information considered to be sensitive because it conveys information about the thoughts and views of individuals.<sup>147</sup> It includes cable television subscription records, cable television viewing history, video rental records, records of library use, and records of reading material purchased. Also in this category are the content of recorded conversations, political opinions, religious or philosophical beliefs, trade union membership, and voting information.<sup>148</sup> While activity on the Internet often reveals similar information,<sup>149</sup> we did not include computer-related identifiers in this category. Those identifiers were included under the computer use category.

A variety of statutes at the federal and state level address television viewing, video rental, and library use.<sup>150</sup> Records of books purchased may be protected as a constitutional right or through state legislation.<sup>151</sup> The

---

PUB. POL'Y 237 (2012) (examining the inconsistent treatment of death records including images in terms of access and privacy); Hu, *supra* note 143, at 1484–91 (explaining how face-recognition technologies and practices are growing while individuals are largely unaware of the privacy threats).

147. See Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387, 408 (2008) (“[W]e should understand intellectual privacy as a series of nested protections, with the most private area of our thoughts at the center, and gradually expanding outward to encompass our reading, our communications, and our expressive dealings with others.”).

148. Federal and state wiretap laws offer protection against undisclosed recording of conversations in some contexts. See, e.g., 18 U.S.C. §§ 2510–2522. Anonymity of speech is given some First Amendment protection, particularly in cases involving speech related to political activity and voting, religious freedom, trade union membership, and other associational activities. See *NAACP v. Alabama ex rel. Patterson*, 357 U.S. 449, 462–63 (1958) (holding that compelled disclosure of names of members would burden the right to freedom of association).

149. See Julie E. Cohen, *A Right to Read Anonymously: A Closer Look at “Copyright Management” in Cyberspace*, 28 CONN. L. REV. 981, 981–82 (1996); Richards, *supra* note 147, at 388–89.

150. The Video Privacy Protection Act, 18 U.S.C. § 2710 (2006), restricts video rental companies from sharing individuals’ viewing habits. The Cable Communications Policy Act, 47 U.S.C. § 551(a)–(h) (2006), restricts disclosure of personally identifiable cable viewing records by cable television companies. Forty-eight states and the District of Columbia have statutes protecting some level of confidentiality of library use, and the remaining states, Hawaii and Kentucky, have Attorney General Opinions stating that the law of the state extends similar protection. The American Library Association maintains links to these state library confidentiality laws. *State Privacy Laws Regarding Library Records*, AM. LIBRARY ASS’N, <http://www.ala.org/advocacy/privacyconfidentiality/privacy/stateprivacy> (last visited Feb. 2, 2016).

151. Bookstores have asserted the confidentiality of books purchased. See *Tattered Cover, Inc. v. City of Thornton*, 44 P.3d 1044, 1053 (Colo. 2002) (finding that the Colorado constitution provided a higher level of protection than the federal constitution

confidentiality of associations that reveal unpopular political beliefs has been recognized as protected under the U.S. Constitution,<sup>152</sup> and scholars have advocated for recognition of other intellectual privacy protections.<sup>153</sup>

### 12. Location

Information in the “location” category includes geolocation information, home address, school address, and work address. Another information type in this category is full zip code with four or more digits that are associated with an identified individual.

At present, geolocation information is considered sensitive information requiring limits on disclosure under the Federal Trade Commission Act’s Section 5 “unfair or deceptive trade practices” protections.<sup>154</sup> Geolocation information is considered sensitive when it is collected from a child using the Internet, and collection of this information is restricted under the Children’s Online Privacy Protection Act.<sup>155</sup> Travel patterns evident in geolocation information are arguably sufficiently sensitive to merit Fourth Amendment recognition.<sup>156</sup>

---

when First Amendment and Fourth Amendment interests intersect in the case of law enforcement seeking book purchase records). California extends protection for e-reader privacy in the Reader Privacy Act, CAL. CIV. CODE § 1798.90 (2011).

152. *NAACP*, 357 U.S. at 463.

153. See Anita L. Allen, *Associational Privacy and the First Amendment: NAACP v. Alabama, Privacy, and Data Protection*, 1 ALA. C.R. & C.L. L. REV. 1 (2011) (reviewing the growth of protections for associational privacy, decisional privacy, and anonymity after *NAACP v. Alabama*); Neil Richards, *INTELLECTUAL PRIVACY* 5, 161 (2015) (examining law, policy, and practical approaches to “safeguard the processes of intellectual explorations and belief formation” and advocating for recognition that “intellectual records are sensitive records that demand higher protection than other kinds of data”). Additionally, the European Union provides protection for “personal data revealing . . . political opinions, religious or philosophical beliefs, [or] trade-union membership . . .” Council Directive 95/46, art. 8, 1995 O.J. (L 281) 38 (EC).

154. *Prepared Statement of the Federal Trade Commission on S. 271, The Location Privacy Protection Act of 2014 Before the S. Comm. on the Judiciary Subcomm. for Privacy, Tech. and the Law*, 113th Cong. (2014) (statement of Jessica L. Rich, Dir. Bureau of Consumer Protection, Fed. Trade Comm’n), [https://www.ftc.gov/system/files/documents/public\\_statements/313671/140604locationprivacyact.pdf](https://www.ftc.gov/system/files/documents/public_statements/313671/140604locationprivacyact.pdf); *FTC Casebook: Goldenshores Technologies, LLC & Eric M Geidl*, INT’L ASS’N PRIVACY PROF’LS, <https://privacyassociation.org/resources/ftc-casebook/goldenshores-technologies-llc-erik-m-geidl> (last visited Apr. 3, 2015) (providing case documents and analysis of FTC action regarding the deceptive use of geolocation data); *FTC Casebook: Aspen Way Enterprises*, INT’L ASS’N PRIVACY PROF’LS, <https://privacyassociation.org/resources/ftc-casebook/aspen-way-enterprises> (last visited Apr. 3, 2015) (discussing geolocation data).

155. See Children’s Online Privacy Protection Act (COPPA), 15 U.S.C. §§ 6501–6506 (2012); 16 C.F.R. §§ 312.1–13 (2015).

156. See Hu, *supra* note 143, at 1481–82, 1500–03 (questioning the capacity of current Fourth Amendment jurisprudence to prevent growth in body and device tracking

### 13. *Sexual Activities*

The “sexual activities” category contains two information types: information about sexual activity and video or audio recordings of an identified individual engaged in a sexual act. Information and images relating to sexual activity are sometimes protected through privacy torts and through state statutes designed to address hidden cameras and non-consensual distribution.<sup>157</sup>

## IV. STUDY DESIGN AND METHODOLOGY

To better understand the privacy interests that might be implicated by public access to court records, we selected a random sample of court records from a large corpus of North Carolina Supreme Court case files that are part of an ongoing digitization project by the UNC Law Library. We coded these documents in order to collect data about the frequency of appearance of sensitive information in the records, as well as other contextual information about the documents and the underlying cases. Once the coding was complete, we used statistical software to analyze the data we collected.

### A. CORPUS OF COURT RECORDS UNDER STUDY

The UNC Law library has approximately 400 bound volumes of North Carolina Supreme Court case filings from 1928 to 2000. In 2013, the library embarked on an ambitious project to digitize some of these records and eventually make a version of them available and searchable online.<sup>158</sup> To date, the library has digitized 12,137 briefs and other filings from the North Carolina Supreme Court comprising 535,106 pages.<sup>159</sup> Each case

---

practices and detailing how geolocation tracking is expanding); Joel R. Reidenberg, *Privacy in Public*, 69 U. MIAMI L. REV. 141, 155–57 (2014) (proposing a new Fourth Amendment doctrine to address geolocational and other types of privacy).

157. Andrew J. McClurg, *Kiss and Tell: Protecting Intimate Relationship Privacy Through Implied Contracts of Confidentiality*, 74 U. CIN. L. REV. 887 (2006); MASS. GEN. LAWS ch. 272, § 105 (2015) (criminalizing the taking of “up-skirt” photos); N.J. STAT. ANN. § 2C:14-9 (West 2015) (criminalizing “revenge porn”).

158. The North Carolina Supreme Court began providing electronic access to its filings in 2000, but does not provide electronic access to records from prior years. Copies of filings prior to 2000 were shared with several non-court libraries in the state, including the UNC Law Library.

159. For the period 1984–2000, the library has scanned and digitized the case filings from 2255 cases heard by the North Carolina Supreme Court; this is approximately 85% of the cases in which the court issued a decision during this time period. For reasons unknown, the court did not send the library any case filings from approximately 15% of

file in this corpus contains at least one merits brief,<sup>160</sup> which may include an appendix. When a brief does have an appendix, it may contain court transcripts, witness testimony, and exhibits entered into evidence such as bank statements, medical records, psychological evaluations, and emails.

Our study used a stratified random sample by year of documents pulled from this corpus of records spanning the time period 1984 to 2000. These documents included briefs and petitions for discretionary review, along with their associated appendices. We did not review the “record on appeal,” which is a separate filing containing, *inter alia*, copies of the case pleadings, jury instructions, transcripts, and other evidence filed in the lower courts.<sup>161</sup> In total, we analyzed 504 documents drawn from 466 cases.<sup>162</sup> One hundred and ninety-eight (39%) of these documents contained an appendix.

## B. CODING AND ANALYSIS

We then performed content analysis on the documents in our sample.<sup>163</sup> This involved coding each document based on its content and case characteristics. The coding, which was performed by a team of eleven research assistants, captured information about each document (e.g., document type, length); information about the underlying case (e.g., date, case type); the type of sensitive information found in the record (e.g., social security number, HIV status); the general category of sensitive information (e.g., financial, health); and information about the location of

---

the cases the court heard and decided during this time period. These digitized records are being redacted in preparation for posting as searchable documents on the Internet.

160. In addition to a brief filed by the appellant, the case files also contain briefs by the appellee, reply briefs, and *amicus curiae* briefs. For cases that do not involve an appeal as of right to the North Carolina Supreme Court, the case file will also contain a petition for discretionary review.

161. See N.C. R. APP. P. 9. The UNC Law Library is not digitizing these records.

162. The number of documents exceeds the number of cases because some cases produced more than one document in our sample.

163. “Content analysis refers to the systematic reading and analysis of texts.” Lee Petherbridge & R. Polk Wagner, *The Federal Circuit and Patentability: An Empirical Assessment of the Law of Obviousness*, 85 TEX. L. REV. 2051, 2070 (2007). In the context of legal scholarship, content analysis is typically performed on judicial opinions. See Mark A. Hall and Ronald F. Wright, *Systematic Content Analysis of Judicial Opinions*, 96 CALIF. L. REV. 63, 64–65 (2008). This methodology is also valuable in the analysis of court records because it allows us to empirically test scholars’ intuitions about the content of court files. See David S. Ardia, *Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity Under Section 230 of the Communications Decency Act*, 43 LOY. L.A. L. REV. 373, 413 (2010) (noting that content analysis as a methodology allows scholars to move beyond anecdotes by generating objective, falsifiable, and reproducible data).

the sensitive information within the document (e.g., brief body, appendix).<sup>164</sup>

The selection and identification of sensitive information types was one of the central challenges of this study. As we described in Part III, there is no single, comprehensive list of private and sensitive information that we could utilize at the start of this project. Existing privacy laws, regulations, and customs have created a patchwork of inconsistent approaches and there is, as yet, no consensus among privacy scholars as to what information should be deemed private or sensitive in the context of court records.<sup>165</sup> Nevertheless, in order to make this study possible, we created a list of 140 sensitive information types based on a survey we conducted of existing legal authority and privacy scholarship and grouped those information types into thirteen categories.<sup>166</sup>

Once the documents were coded, we used STATA, a general-purpose statistical software package, to determine the frequency of appearance of each sensitive information type and to identify relationships, patterns, and correlations between different information types and other coded variables, including trends over time. A summary of our findings is included in Part V.

To check the reliability of the coding process we conducted two phases of testing.<sup>167</sup> First, we “pilot tested” a preliminary version of our coding form by having our coders review an identical set of five case documents.<sup>168</sup> We also reviewed those documents ourselves and compared the results of all coders. This resulted in minor alterations to the coding scheme and coder instructions. Second, we conducted a formal test of reliability at the conclusion of the coding process by selecting a random sample of fifty documents from the 504 documents in the study set.<sup>169</sup> We assigned each

---

164. The coders used Qualtrics, an online survey platform, to record their observations. The coding instrument and codebook are available on the authors' website. *See Media Law Resources*, UNC CTR. FOR MEDIA L. & POL'Y, <http://medialaw.unc.edu/resources> (last visited Feb. 1, 2016).

165. *See* Conley et al., *supra* note 13, at 775 (concluding that a complex body of rules, regulations, principles, and policies govern the creation of court records and access to them).

166. *See supra* Part III.

167. *See* Petherbridge & Wagner, *supra* note 163, at 2074 (noting that reliability testing is crucial because “the process of content analysis . . . is inherently subject to some level of subjectivity”).

168. *See* LEE EPSTEIN & ANDREW D. MARTIN, AN INTRODUCTION TO EMPIRICAL LEGAL RESEARCH 101 (2014) (suggesting the use of a “pilot study” to pretest content coding schemas).

169. There is no bright-line standard dictating the sample size to be used when doing reliability testing. *See* Petherbridge & Wagner, *supra* note 163, at 2074, n.118 (stating

of the documents in this subset to a coder who had not previously coded the document. We then compared the results of the two codings in order to assess the degree of inter-coder reliability.<sup>170</sup>

## V. RESULTS AND DISCUSSION

We begin in this part by presenting descriptive statistical information about the court records in our sample and the sensitive information they contain. We then examine the extent to which different types of sensitive information are related to various case and document characteristics.

### A. DESCRIPTIVE STATISTICS

#### 1. *Sample Summary*

The 504 court records that we reviewed contained a total of 24,156 pages, with a mean document length of 47.9 pages.<sup>171</sup> Table 1 presents a breakdown of the various document types that were in our sample, including the number of documents with an appendix and the length (in pages) for each document type.<sup>172</sup> As shown in Table 1, 198 (39%) of these documents included an appendix. Not surprisingly, documents that contained an appendix were substantially longer (mean length of seventy-one pages) than documents without an appendix (mean length of thirty-three pages).<sup>173</sup> We also found considerable variability among the

---

that researchers suggest that at least a ten-percent sample be used) (citing Stephen Lacy & Daniel Riffe, *Sampling Error and Selecting Intercoder Reliability Samples for Nominal Content Categories*, 73 JOURNALISM & MASS COMM. Q. 963, 969–73 (1996)).

170. The percentage rate of agreement and “Krippendorff’s alpha,” see KLAUS KRIPPENDORFF, *CONTENT ANALYSIS: AN INTRODUCTION TO ITS METHODOLOGY* 221–30 (2d ed. 2004), for each of the variables is listed on the coding form, which is available on the authors’ website. See *Media Law Resources*, UNC CTR. FOR MEDIA L. & POL’Y, <http://medialaw.unc.edu/resources> (last visited Feb. 1, 2016).

171. The median document length was thirty-two pages. We report the median length in addition to the mean because document length was not normally distributed within the sample. The median document length is therefore a better measure of central tendency.

172. The document types were coded based on the document title on the first page of the brief or petition. According to the North Carolina Rules of Appellate Procedure, “The Title of the Document should reflect the position of the filing party both at the trial level and on the appeal, e.g., DEFENDANT-APPELLANT’S BRIEF, PLAINTIFF-APPELLEE’S BRIEF, or BRIEF FOR THE STATE.” N.C. R. APP. P. app. E (1975). “Briefs for the State” are briefs filed by the State of North Carolina as either an appellant or appellee; the brief captions do not designate the specific role of the State.

173. These results suggest that there is a statistically significant difference between the distributions of page lengths for documents with and without an appendix ( $z = -9.372$ ,  $p = 0.0000$ ). We utilized the Wilcoxon-Mann-Whitney test for statistical

document types with regard to the inclusion of an appendix. Nearly all “Petitions for Discretionary Review” contained an appendix (98%)<sup>174</sup> while “Briefs for the State” were the least likely document type to include an appendix (16%).<sup>175</sup>

The most commonly occurring document type in the sample was briefs filed by the appellant, which constituted almost half of the documents (41%). The sample also included a number of non-party *amicus curiae* briefs (3%), which tended to be the longest documents in the sample. Although our sample did not include any reply briefs, we know from our review of the North Carolina Supreme Court’s case files that a small number of reply briefs also exist in the population under study.<sup>176</sup>

---

significance because document length was not normally distributed within the sample. We return to the importance of the appendices in Part B.

174. The high proportion of petitions that included an appendix is likely due to the requirements in the North Carolina Rules of Appellate Procedure, which state that a petition for discretionary review “shall be accompanied by a copy of the opinion of the Court of Appeals when filed after determination by that court.” N.C. R. APP. P. 15(c). We did not code for the type of documents attached as appendices, so we cannot state what proportion of the appendices included only a copy of the lower court’s opinion.

175. The difference between these document types with regard to their inclusion of appendices is statistically significant (chi-square with five degrees of freedom = 97.9666,  $p = 0.000$ ).

176. As with all random sampling approaches, there is a chance that our sample of documents did not capture the entire range of characteristics in the population of North Carolina Supreme Court records. It is likely, however, that any characteristics that are not in the sample appear very infrequently in the target population. Prior to 2009, the North Carolina Rules of Appellate Procedure did not permit the filing of a reply brief unless the court requested such a brief or certain special circumstances existed. *See* N.C. R. APP. P. 28(h) (1975).

Table 1: Frequency of document types in the sample, including the number of documents with an appendix [n, (%)] and length (in pages) for each document type.

Document Type	n	Appendix	Page Length	
			Mean	Median
Brief of Appellant	212	87 (41%)	54.4	34
Brief of Appellee	140	41 (29%)	37.8	30
Brief for the State	87	14 (16%)	46.0	30
Petition for Discretionary Review	46	45 (98%)	46.7	36
Brief of <i>Amicus Curiae</i>	16	9 (56%)	67.6	36
Other <sup>177</sup>	3	2 (67%)	30.7	26
All Document Types	504	198 (39%)	47.9	32

Our sample of documents came from cases decided by the North Carolina Supreme Court between 1984 and 2000, the years immediately preceding the introduction of electronic filing in North Carolina when additional rules regarding the redaction of sensitive information took effect.<sup>178</sup> Table 2 lists the number of documents in the sample by the type of case and the year in which the court issued its decision in the case. Because we selected a random stratified sample by year, the totals for each year are relatively constant, with a spike in the number of documents selected from cases decided in 1986 and a drop in documents from 2000. Although there was some variation in the proportions each year, nearly two-thirds of the documents in the sample came from civil cases (62%), slightly more than a third came from criminal cases (36%), and only a small proportion (1%) came from juvenile proceedings.<sup>179</sup>

177. The sample also included a motion to amend, guardian ad litem's brief, and brief by a cross-appellant.

178. See *supra* note 10 and citations therein.

179. Each state has special courts—typically called juvenile courts—that have jurisdiction over cases involving children under a specified age. See *Juvenile Court*, BLACK'S LAW DICTIONARY (10th ed. 2014). Juvenile court proceedings are civil as opposed to criminal. *Id.* In North Carolina, “[a] person who has not reached the person’s eighteenth birthday and is not married, emancipated, or a member of the Armed Forces of the United States” is eligible for juvenile court if the case relates to abuse, neglect, or dependency. N.C. GEN. STAT. § 7B-101(14). Persons sixteen years and older who are charged with certain criminal violations or infractions are not eligible for juvenile court in North Carolina. See N.C. GEN. STAT. § 7B-1501(7) & (27). In our coding of the juvenile cases, we did not differentiate between delinquency cases and abuse, neglect, and dependency cases.

Table 2: Number of documents by case type and year of North Carolina Supreme Court's decision, including overall percentages for each case type.

Year	Case Type			Total
	Civil	Criminal	Juvenile	
1984	9	11	1	21
1985	16	12	0	28
1986	20	21	0	41
1987	23	11	1	35
1988	15	15	1	31
1989	11	12	0	23
1990	19	11	1	31
1991	24	10	0	34
1992	23	10	1	34
1993	20	9	0	29
1994	20	17	0	37
1995	23	9	0	32
1996	23	11	0	34
1997	20	10	1	31
1998	21	6	0	27
1999	18	5	1	24
2000	8	4	0	12
Total	313 (62%)	184 (36%)	7 (1%)	504

The cases themselves covered a wide variety of subject areas, ranging from appeals challenging death penalty sentences to workers' compensation determinations. To facilitate the coding of case subject areas, we adopted the twelve appellate case type designations created by the Court Statistics Project at the National Center for State Courts.<sup>180</sup> Not surprisingly, given that nearly two-thirds of the documents in our sample came from civil cases, the most commonly occurring appellate subject area

180. The appellate case types are: (1) Death Penalty; (2) Felony (non-Death Penalty); (3) Misdemeanor; (4) Criminal-Other; (5) Tort, Contract, and Real Property; (6) Probate; (7) Family; (8) Juvenile; (9) Civil-Other; (10) Workers' Compensation; (11) Revenue (Tax); and (12) Administrative Agency-Other. *See* COURT STATISTICS PROJECT, STATE COURT GUIDE TO STATISTICAL REPORTING 39–44 (2014), <http://www.courtstatistics.org/~media/Microsites/Files/CSP/State%20Court%20Guide%20to%20Statistical%20Reporting%20v%202011.pdf>. The Court Statistics Project at the National Center for State Courts created these categories in order to provide a “standardized reporting framework for state court caseload statistics designed to promote intelligent comparisons among state courts.” *Id.* at 1.

was “tort, contract, and real property,” which constituted 32% of the documents in our sample. The next most common subject area was “felony (non-death penalty),” which arose in 26% of the documents, followed by appeals of administrative agency decisions, which appeared in nearly 9% of the documents in the sample. Documents from death penalty cases constituted 7% of the sample, yet they contained more than a quarter (28%) of the sensitive information we found.<sup>181</sup>

## 2. *Sensitive Information Summary*

Although a wide variety of sensitive information appears in the court records we sampled, it is not uniformly distributed throughout the records. Most of the documents contained relatively few incidences of sensitive information while a handful of documents contained a large number of pieces of sensitive information. Figure 1 presents a histogram of the frequency of sensitive information per document. It shows a pronounced rightward skew indicating that sensitive information is not “normally” distributed throughout the records.<sup>182</sup> In other words, the histogram is asymmetrical and does not have the classic bell shaped curve that would indicate that most documents fall within the middle of the range. Instead, the vast majority of documents contained fewer than forty pieces of sensitive information while only a few documents contained more than 400 pieces of sensitive information. At the far right of the graph we see that several documents contained more than 1,000 pieces of sensitive information. Overall, the records we reviewed contained an average of 113 appearances of sensitive information per document, with a median of thirty-six appearances of sensitive information.<sup>183</sup>

We saw considerable variation in the frequency of sensitive information among the different document and case types. Table 3 presents the median frequency of sensitive information by document type along with the location (brief body or appendix) where the information appeared.<sup>184</sup> Figure 2 presents similar information by case type.

---

181. We discuss the potential implications of this finding in Section VI.A.3.

182. A rightward skew is when the long tail is on the right side of the peak, which is also called a positive skew.

183. The standard deviation for the frequency of sensitive information coded per document is 209.07 and the interquartile range, covering the middle 50% of the observed frequencies, is 11–122.

184. The difference in the frequency of sensitive information between brief bodies and appendices is statistically significant (paired *t*-test with 104 degrees of freedom = -3.5484, *p* = 0.0006). We report the median frequency in Tables 3 and 4, rather than the mean, because the frequency of sensitive information was not normally distributed.

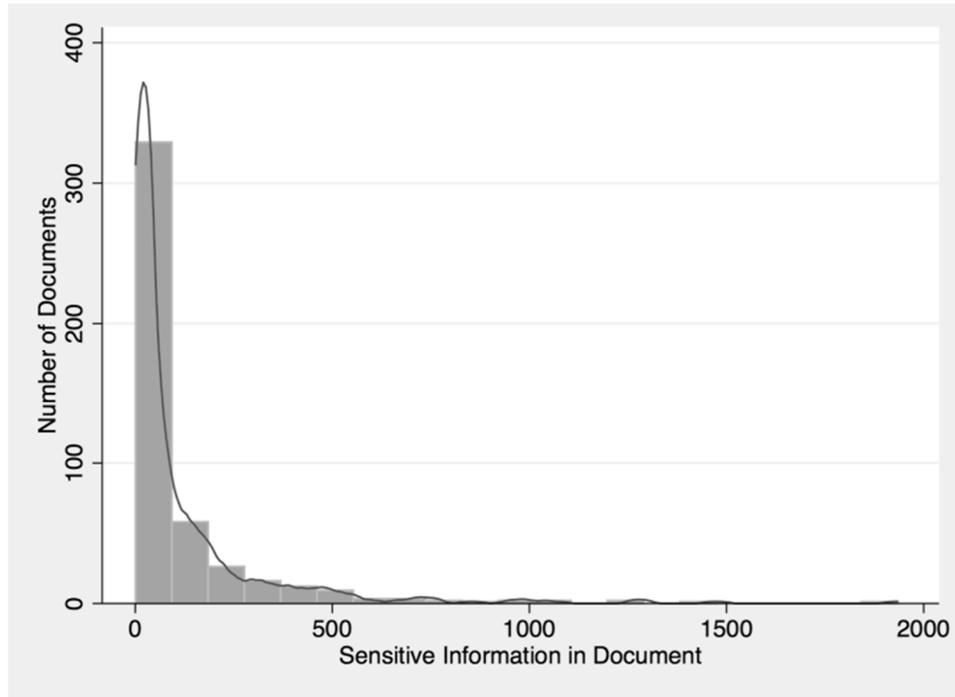


Figure 1: Histogram of frequency of sensitive information per document overlaid with kernel density plot.

Table 3: Median frequency of sensitive information coded per document, listed by document type and location within the document.

Document Type	Median Frequency of Sensitive Information		
	Brief Body	Appendix	Overall
Brief for the State	96.0	30.0	105.5
Brief of Appellant	39.0	24.0	42.0
Petition for Discretionary Review	14.5	17.0	22.0
Brief of Appellee	12.5	10.5	12.0
Brief of <i>Amicus Curiae</i>	8.0	215.5	10.5
Other	6.5	3.0	8.0
All Document Types	29.0	19.0	36.0

Figure 2 presents similar information by case type and reveals that criminal cases had substantially more sensitive information per document than either civil or juvenile cases. In fact, the median frequency of sensitive information in documents filed in criminal cases was approximately five times that of documents filed in either civil or juvenile cases. Figure 2 also reveals that in criminal and juvenile cases, sensitive information appeared

much more frequently in the brief body than in the appendix. In civil cases, sensitive information appeared with equal frequency in both appendices and briefs. We return to the role of appendices in Section V.B.

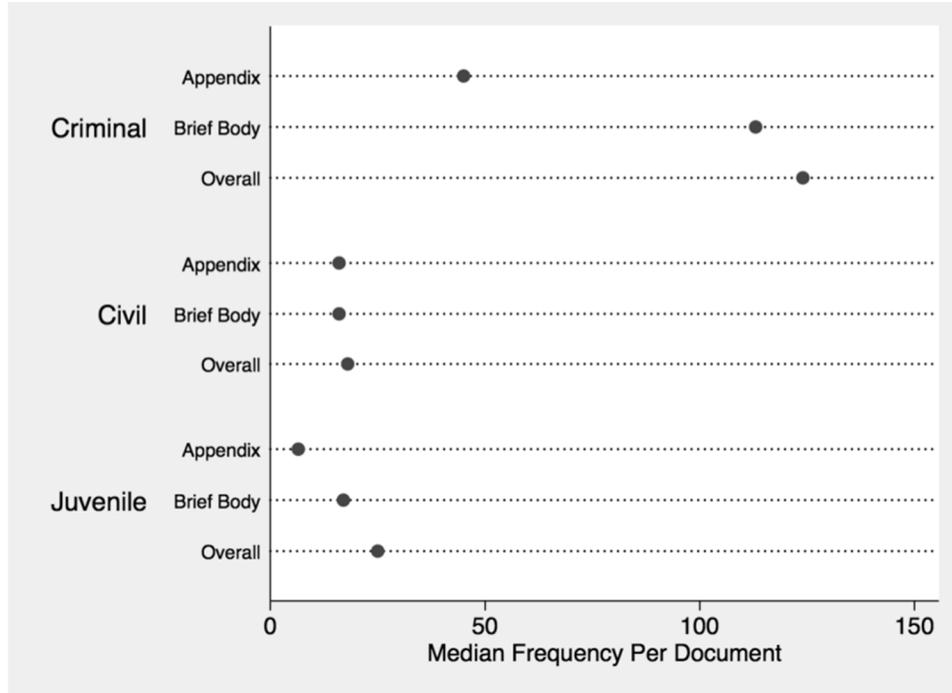


Figure 2: Dot plot of median frequency of sensitive information per document, by case type and location within the document.

As we noted in Part III, we grouped the specific sensitive information types into thirteen categories in order to facilitate comparisons between sensitive information types that shared similar characteristics. Table 4 reports the number of documents in the sample that contained sensitive information falling within each of these information categories. As Table 4 shows, information types in the “location” category appeared in more documents than any other category, appearing in 67% of the documents in the sample. Information in the “identity” and “criminal proceedings” categories also appeared in more than half of the documents, occurring in 66% and 56% of the documents, respectively. Overall, information in seven of the thirteen categories appeared in at least 20% of the documents.<sup>185</sup> Information in each of the remaining six categories appeared in fewer than 8% of the documents.

185. We break down the specific information types within these categories in Section V.B.

Table 4: Number of documents that contained sensitive information, listed by category of sensitive information, including the percentage of the total sample size ( $n = 504$ ) and median number of times per document that information in that category appeared.

Information Category	Documents in Sample		Median
	$n$	%	Per Doc
Location	336	67%	4
Identity	331	66%	3
Criminal Proceedings	280	56%	54.5
Health	205	41%	5
Assets	175	35%	6
Financial Information	134	27%	4
Civil Proceedings	103	20%	4
No sensitive information	37	7%	-
Employment	33	7%	4
Sexual Activities	31	6%	6
Intellectual Pursuits	23	5%	4
Education	6	1%	3.5
Images	1	0%	17
Computer Use	0	0%	-

A few categories stand out in Table 4 because of their relative absence in the documents. Information about “sexual activities” appeared infrequently, as did information in the “intellectual pursuits” category, which includes religious beliefs, political opinions, and voting and reading records. Information about “education” was also mostly absent from the documents as were photos and videos captured by the “images” category. None of the documents contained any sensitive information in the “computer use” category (e.g., user names, passwords, and search history). Moreover, thirty-seven documents in the sample (7%) did not contain any of the sensitive information types that we coded for in this project.<sup>186</sup>

Table 4 also presents the median number of times per document that information in each category appeared. For most of the information categories, sensitive information appeared between three and six times per document. There are two outliers, however. Information in the “criminal proceedings” category appeared far more frequently in the documents than any other category (median appearance per document of 54.5), showing up approximately nine to eighteen times as often as information in the other categories. The “images” category was the other outlier, with a median of

---

186. A full list of the information types we coded is included in the Appendix.

seventeen appearances of sensitive information per document.<sup>187</sup> Although sensitive information in the “images” and “sexual activities” categories did not appear in very many documents, when they did appear, they generally did so with greater frequency than information in all of the other categories excluding “criminal proceedings.”

## B. ANALYSIS

It would be of little value to the debate over privacy and public access to simply add up the total number of times that various types of sensitive information appeared in court records. Indeed, we knew going into this study that court records are replete with sensitive information. Instead, in reporting and interpreting the results, we focus on the relative differences between document types, case types, and information categories rather than on the absolute numbers.

As noted in Part III and discussed more fully below, we purposefully coded a broad range of sensitive information types. Not all of the information that we identified presents the same privacy concerns. In the following sections we discuss the types of sensitive information that we found in the documents and examine the context in which they appeared.

### 1. *Variations Within and Among Information Categories*

Not surprisingly, the court records did not contain every category or type of sensitive information in equal measure. As Table 4 shows, information relating to location, identity, criminal proceedings, health, assets, finances, and civil proceedings appeared in many more documents than information that falls within the remaining six categories. We observed the same “top seven” categories of information when we calculated the total frequency of sensitive information throughout all of the records, but in a slightly different order.<sup>188</sup> For example, although the criminal proceedings category was only the third most frequently occurring information category on a per document basis (information types in this category appeared in 56% of the documents) it far exceeded every other category of information on the basis of total frequency of

---

187. This may be due to the fact that only one document in the sample contained information that fell within this category. A sub-sample consisting of only a single observation is too small to be statistically significant.

188. The top seven categories in terms of total appearance of sensitive information were criminal proceedings ( $n = 38,136$ ), health (3,549), identity (3,217), assets (2,385), location (2,128), civil proceedings (1,428), and financial information (1,097).

appearance.<sup>189</sup> In other words, information related to criminal proceedings not only appeared in most court records, it also appeared more often in those records than any other category of sensitive information.

We might speculate that information related to criminal proceedings appears more frequently because criminal cases may be more common than civil cases. Documents from criminal cases, however, made up only 36% of the sample,<sup>190</sup> so the higher frequency of criminal information is not due to a larger number of criminal documents. Instead, criminal information was dispersed across all of the document types and case types. It is not just criminal cases that contain criminal information; this information appeared in a wide variety of contexts. We consider this further in the next section.

Turning to the individual information types in each of the most frequently occurring information categories, we saw a general pattern in the distribution of sensitive information. Figure 3 presents dot plots for the eight most frequently occurring information categories.<sup>191</sup> In each category, a few information types appeared far more often than the other information types in that category. This pattern was most evident for the financial information category, where information about an individual's compensation far outnumbered the other types of financial information in terms of frequency of appearance in court records.<sup>192</sup> This pattern was less pronounced for the assets and location categories, which had three and four types respectively of sensitive information that constituted more than 10% of their category's total. For the criminal proceedings and civil proceedings categories, the distribution was also more evenly spread; both of these categories had at least three information types that comprised 10% or more of their category's total.<sup>193</sup>

---

189. Information in the criminal proceedings category appeared 38,136 times in the sample. Information in the next highest category, health, had an overall frequency of appearance of 3,549. The substantially higher number of median appearances per document of information related to criminal proceedings as shown in Table 4 suggests this disparity as well.

190. See Table 2.

191. We present dot plots for the eight most frequently occurring categories, rather than just the top seven categories, to make full use of the space available in Figure 3. Note that the horizontal axes for these plots varied from a maximum frequency of 250 (for the "employment" category) to 15,000 (for the "criminal proceedings" category). The axes were presented in this way in order to allow clearer comparisons of frequency *within* each category.

192. There were surprisingly few incidences of bank account numbers ( $n = 9$ ), credit card numbers (2), or other financial account numbers (4) in the sample. See Figure 3.

193. The criminal proceedings category dwarfed all other categories in terms of overall frequency of appearance. Unlike the other categories, six information types in the

As Figure 3 confirms, the criminal proceedings category far exceeded every other category of information on the basis of total frequency of sensitive information in the court records. Names of witnesses in criminal cases appeared more often than any other coded information type, followed somewhat distantly by the name of an individual who was the subject of a criminal investigation. The name of a victim of criminal activity other than rape (rape victim name is a separate information type) was the third most frequently occurring information type in the criminal proceedings category—and the third most frequently occurring information type overall. It is not until after the fifth most frequently occurring criminal information type, “conviction,” that information in the other categories begin to place in the rankings of most frequently occurring sensitive information types.

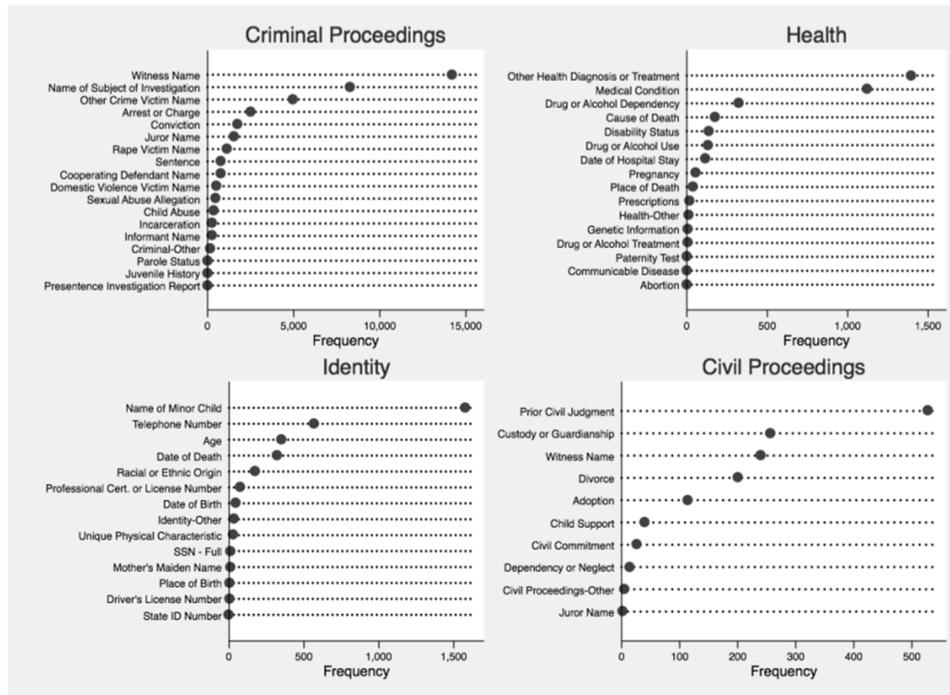


Figure 3A: Frequency of individual information types in the most commonly occurring categories of sensitive information.

criminal proceedings category appeared more than 1,000 times and three types of sensitive information appeared more than 4,900 times. See Figure 3.

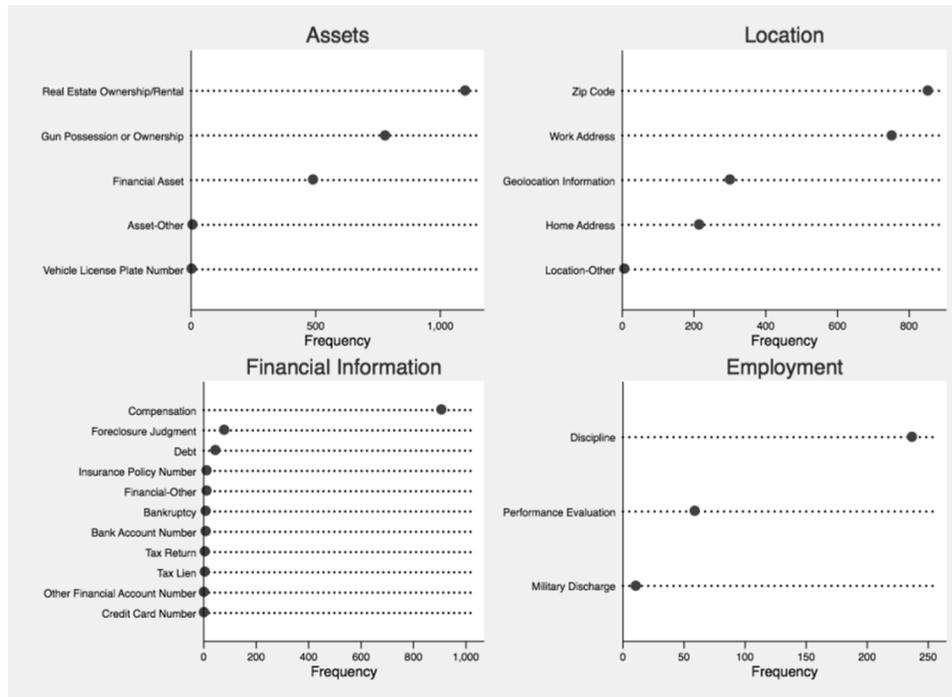


Figure 3B: Frequency of individual information types in the most commonly occurring categories of sensitive information.

The higher frequency of information related to criminal proceedings could be due to the fact that documents filed in criminal cases were, on average, longer than documents filed in other types of cases.<sup>194</sup> We would naturally expect longer documents to have more sensitive information. The data support this intuition, although the relationship between document length and frequency of sensitive information only partially explains the variations in the documents. Figure 4 presents a scatterplot of total sensitive information per document as a function of document length (in pages). The line through the scatterplot is the best-fitting linear regression line that provides an estimate of the relationship between the frequency of sensitive information in a document and the document's length.

194. Documents associated with criminal cases were on average 11.6 pages longer than civil cases: criminal cases had a mean [median] document length of 55.5 [36] pages compared to 43.9 [30] pages for documents filed in civil cases. Juvenile cases had on average the shortest documents, with a mean [median] of 32.3 [36] pages.

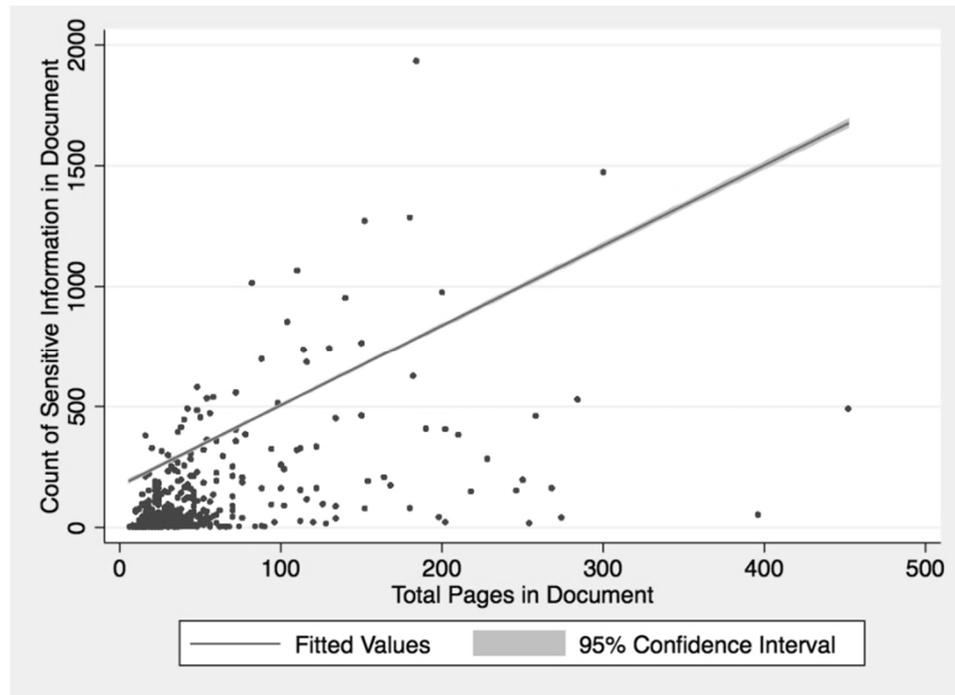


Figure 4: Scatterplot of frequency of sensitive information per document on document length (in pages) with linear regression line.

We draw several conclusions from Figure 4. First, the relationship between total frequency of sensitive information in a document and document length is positive (as documents get longer, we can expect to find more sensitive information). Second, the overall ratio is approximately 1:3 (for each additional page in length, we can expect to find approximately three more pieces of sensitive information).<sup>195</sup> Although as Figure 4 demonstrates, document length is an important indicator of the frequency of sensitive information in a court record, it accounts for only an estimated 33% of the variation in total frequency of sensitive information in the documents.<sup>196</sup> In other words, other independent variables, either alone or in combination, are likely to have a

195. Utilizing ordinary least squares, the regression model's coefficient for page length was 3.363195 ( $n = 52,998$ , std. err. = 0.0205702,  $R^2 = 0.3299$ ,  $p$ -value = 0.000).

196. The linear regression model used in Figure 4 produces an estimate, known as the coefficient of determination ( $R^2$ ), of the fit between the model's prediction of the number of appearances of sensitive information in a document as a function of page length and the actual frequency of sensitive information. For Figure 4,  $R^2$  was 0.3299. This estimate tells us the percentage of the variance in the frequency of sensitive information explained by the model is 32.99%.

more substantial effect than page length on the frequency of sensitive information in a court record.<sup>197</sup>

Indeed, there are signs that other factors are at work when we look at scatterplots comparing the frequency of sensitive information as a function of document length across the three different case types. As Figure 5 shows, criminal cases had a higher density of sensitive information per page than either civil or juvenile cases. As page length increased, the number of pieces of sensitive information in criminal cases increased at a higher rate than it did in civil and juvenile cases. For criminal cases, the ratio between page length and frequency of sensitive information was roughly 1:4.<sup>198</sup> For civil cases, the ratio was approximately 1:1.<sup>199</sup>

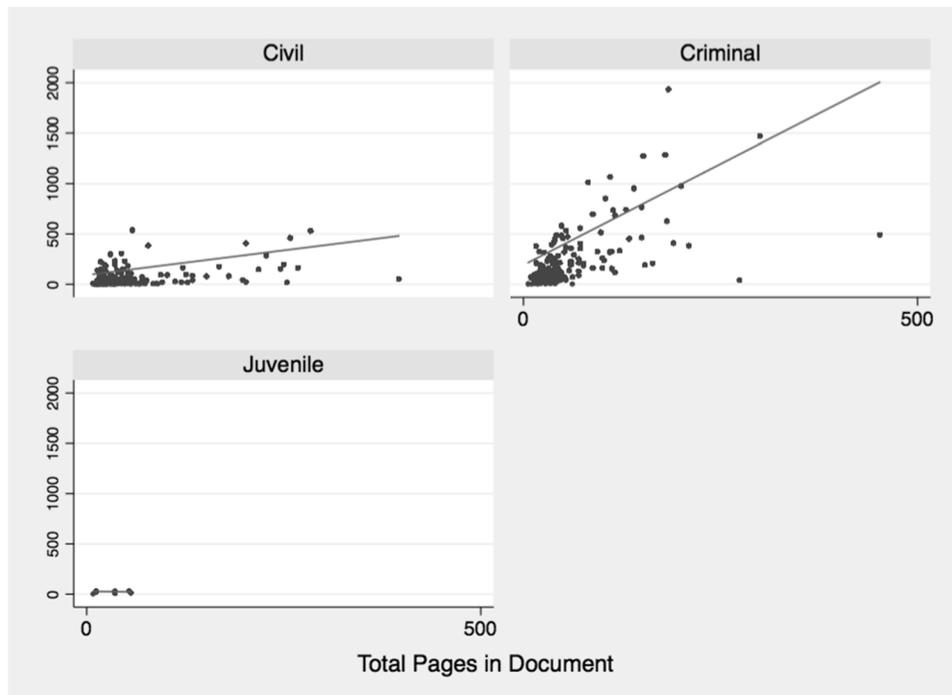


Figure 5: Scatterplot of frequency of sensitive information per document on document length (in pages) by case type with linear regression lines.

197. We report on the results of our multiple regression model in Part V.B.4.

198. Utilizing ordinary least squares, the regression model's coefficient for page length in criminal cases is 4.012137 ( $n = 40,440$ , std. err. = 0.0235338,  $R^2 = 0.4182$ ,  $p$ -value = 0.000).

199. Utilizing ordinary least squares, the regression model's coefficient for page length in civil cases is 0.9802669 ( $n = 12,423$ , std. err. = 0.014221,  $R^2 = 0.2767$ ,  $p$ -value = 0.000). There were too few documents from juvenile cases ( $n = 7$ ) to draw any conclusions about the relationship between document length and frequency of sensitive information.

## 2. *Contextual Variations*

We turn now to the question of whether certain contextual factors influence—or at least are correlated with—the types of sensitive information found in court records. We coded for a number of case and document characteristics that might be linked to the appearance of sensitive information, including case type, case subject area, document type, subject of the information (adult or minor), location of the information (brief body or appendix), and year of case decision.<sup>200</sup> As the preceding discussion noted, we have already seen some variability in the types and extent of sensitive information associated with criminal cases, so we will start by analyzing the role that case type plays in the appearance of sensitive information.

### a) Case Types

We know from Figures 2 and 5 that criminal cases have, on average, more sensitive information than civil and juvenile cases, but we cannot tell from those figures which types of sensitive information are more prevalent in criminal cases. Figure 6 presents the percentage of sensitive information in civil and criminal cases by category of sensitive information.<sup>201</sup> From Figure 6 we can discern some important differences about the extent of sensitive information in civil and criminal cases.

First, sensitive information is not uniformly distributed in all types of cases. The top bar in Figure 6 shows that overall, approximately 75% of the sensitive information we identified appeared in documents filed in criminal cases. Many of the information categories, however, deviated substantially from this 75/25 split.

In civil cases, we found a significantly higher proportion of sensitive information in the assets, civil proceedings, employment, financial, and location categories. In fact, sensitive information in the employment and financial categories appeared almost entirely in civil cases (92% and 94% of the time respectively). Sensitive information in the health, identity, and intellectual pursuits categories, on the other hand, appeared more frequently in documents associated with criminal cases, and information in the education and images categories appeared only in criminal cases. Only

---

200. Other contextual factors may also be relevant, but our focus here is on the case and document characteristics that courts themselves use in their filing systems.

201. Figure 6 does not include documents from juvenile cases because they were too few in number to warrant graphing.

information in the health and sexual activities categories appeared in roughly equal measure in both civil and criminal cases.

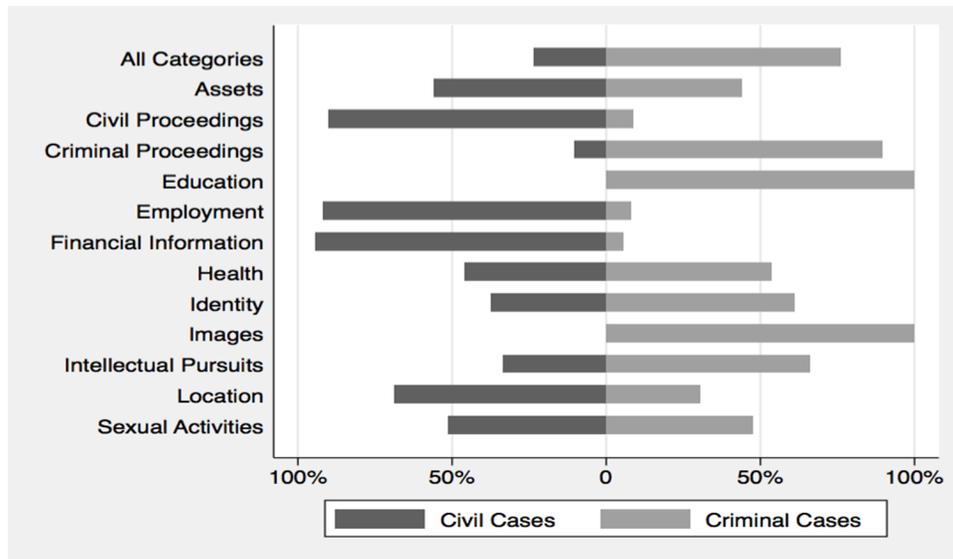


Figure 6: Horizontal bar graph showing percentage of sensitive information in civil and criminal cases by category of sensitive information.

Second, not only did criminal cases evidence more sensitive information than civil and juvenile cases, they also contained a greater variety of sensitive information. Criminal cases contained sensitive information from all of the categories we found in the documents,<sup>202</sup> whereas information from two categories, education and images, were absent from civil cases.<sup>203</sup> We also found that overall, criminal cases contained more types of sensitive information. Of the 140 sensitive information types we coded for in the records, ninety-five distinct types actually appeared in the documents. Although some types appeared exclusively, or nearly so, in civil cases, documents filed in criminal cases contained a greater variety of sensitive information types. Table 5 lists the information types that appeared in documents associated with only one

202. We did not find any information in the documents that fell within the “computer use” category. *See supra* Part III.B.3 (describing the information types in this category).

203. Six of the thirteen information categories were absent from documents filed in juvenile cases.

case type at least 90% of the time.<sup>204</sup> As Table 5 shows, criminal cases had a greater variety of sensitive information types than civil cases.

**Table 5: Information types that appeared in documents associated with only one case type at least 90% of the time.**

Civil Cases	Criminal Cases
Adoption	Abortion
Bank Account Number	Arrest or Charge
Bankruptcy	Child Abuse
Cable Television Subscription Record	Communicable Disease
Child Support	Content of Recorded Conversations
Compensation	Conviction
Debt	Credit Card Number
Discipline	Drug or Alcohol Treatment
Driver's License Number	Full-Face Photograph
Foreclosure Judgment	Genetic Information
Insurance Policy Number	Gun Possession or Ownership
Paternity Test	Incarceration
Performance Evaluation	Informant Name
Political Opinion	Juror Name
Prior Civil Judgment	Juvenile Court History
Professional Cert. or License Number	Military Discharge
SSN - Full	Name of Subject of Investigation
State ID Number	Parole Status
Tax Lien	Photos or Videos of Violence, Abuse or Death
Tax Return	Presentence Investigation Report
Voting Record	Rape Victim Name
	Sentence
	Sexual Abuse Allegation
	Student Discipline
	Student Grades or Performance Evaluation
	Vehicle License Plate Number
	Video Rental Records

204. None of the sensitive information types we coded for appeared more than 10% of the time in juvenile cases. There were several information types, however, that were disproportionately common in juvenile cases: "Adoption," "Custody or Guardianship," "Date of Birth," "Juvenile Court History," "Name of Minor Child," "Pregnancy," and "Sex Life."

## b) Adults and Minors

Most of the sensitive information we found was associated with adults. Overall, only 7% of the sensitive information we coded was associated with an identified minor,<sup>205</sup> and the difference between civil and criminal cases with regard to sensitive information associated with minors was modest (information about minors appeared 6.1% and 7.2% of the time respectively). The percentage of sensitive information about minors was significantly higher in juvenile cases, where 40% of the information that we coded was associated with an identified minor.

Although appeals from juvenile cases are now subject to additional privacy protections under the North Carolina Rules of Appellate Procedure,<sup>206</sup> these protections were not in place during the time period we studied and we found a considerable amount of sensitive information in juvenile cases that was associated with an identified minor. Documents in juvenile cases contained an average of 10.27 pieces of sensitive information connected to an identified minor, with the following information types being the most common: “Name of Minor Child,”<sup>207</sup> “Rape Victim Name,” “Adoption,” “Age,” “Arrest or Charge,” and “Other Health Diagnosis or Treatment.”<sup>208</sup>

We also found a few interesting differences between the information categories with regard to minors across all of the case types. Relative to their baseline percentages, minors were less likely to be associated with

---

205. For our purposes, an identified minor was any individual under the age of eighteen years, regardless of whether he or she met the requirements for juvenile court jurisdiction under North Carolina law. *See supra* note 179 (describing the jurisdictional requirements for juvenile court in North Carolina).

206. In 2006, the North Carolina Rules of Appellate Procedure were amended to provide additional privacy protections for juveniles. The current rules state that “covered juveniles . . . shall be referenced only by the use of initials or pseudonyms in briefs, petitions, and all other filings, and shall be similarly redacted from all documents, exhibits, appendixes, or arguments submitted with such filings” and that a “juvenile’s address and social security number shall be excluded from all filings, documents, exhibits, or arguments with the exception of sealed verbatim transcripts.” N.C. R. APP. P. 3.1(b).

207. As we noted in Part III.B, we coded for “Name of Minor Child” as a specific information type; this information type made up nearly 3% of the total frequency of all sensitive information in the records.

208. We did not record whether the minor in question was a “covered juvenile” under North Carolina law, so we cannot state whether the information we found would violate North Carolina Rule of Appellate Procedure 3.1(b). We can state, however, that we found no appearances in juvenile cases of addresses or social security numbers associated with minors. As we note in Part VI.A.4, the number of documents in our sample from juvenile cases was relatively small ( $n = 7$ ), so we recommend further research on the privacy risks associated with minors.

information related to criminal proceedings (3.7%) and more likely to be associated with information in the education (33.3%), health (10.5%), identity (12.9%), and sexual activities (16.0%) categories.

There also were intriguing variations in the relative proportions of some of the specific information types with regard to minors. Figure 7 shows the percentage of sensitive information associated with adults and minors for the seventeen information types that were identified with minors more than 10% of the time. As Figure 7 reveals, a number of information types were disproportionately associated with minors (i.e., their association with minors was substantially greater than would have been expected based on the overall frequency of sensitive information associated with minors). Information about communicable diseases and minor names, for example, were exclusively identified with minors, and seven of the seventeen information types appeared more than 30% of the time in relation to a minor.

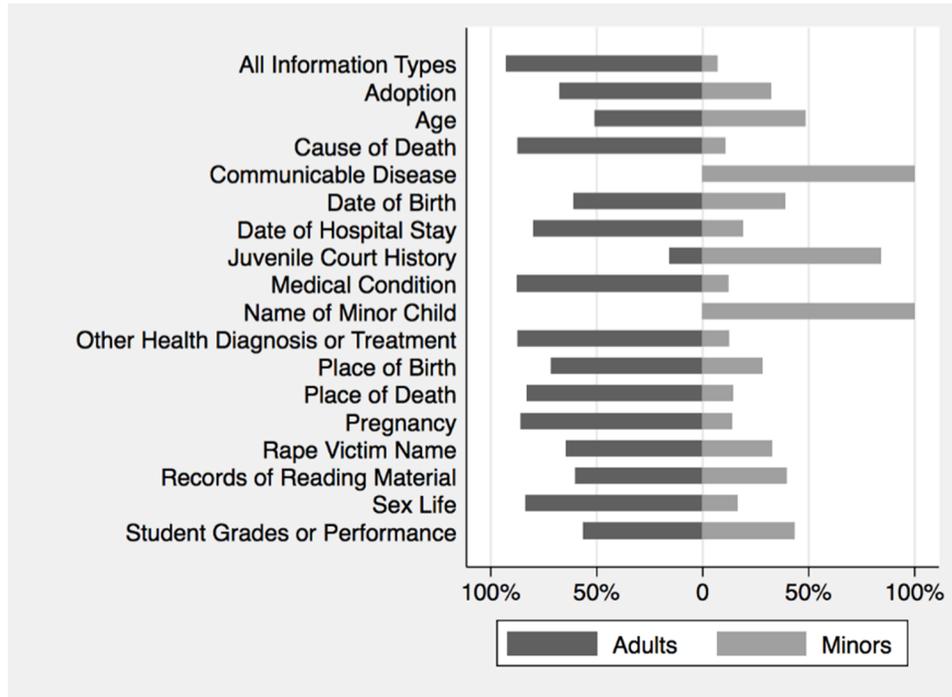


Figure 7: Horizontal bar graph showing percentage of sensitive information types associated with adults and minors. Only information types associated with minors more than 10% of the time are listed.

## c) Appendices

Some scholars and archivists have suggested that appendices included in court records contain more sensitive information than legal briefs,<sup>209</sup> but our data did not bear this out. As we reported in Section V.A, we found that overall, brief bodies contained a higher frequency of sensitive information than appendices.<sup>210</sup> As the dot plot in Figure 2 showed, this disparity was particularly evident in criminal and juvenile cases; for civil cases, sensitive information appeared with equal frequency in both the appendices and brief bodies.<sup>211</sup>

Several information types, however, were more prevalent in appendices. Figure 8 lists the information types that appeared more than 30% of the time in appendices. Of the ninety-five information types we identified in the documents, sixteen appeared more than 30% of the time in an appendix, a marked deviation from the overall proportion of sensitive information in appendices (14%) as shown in the top bar in Figure 8. Moreover, seven information types appeared more than 50% of the time in an appendix, and three types appeared only in the appendices: “SSN - Full,” “State ID Number,” and “Video Rental Records.”

As Figure 8 shows, only seven of the ninety-five information types that we identified in the records appeared more often in appendices. On the other hand, twenty-seven information types appeared exclusively in the brief bodies, including: “Abortion,” “Adoption,” “Bankruptcy,” “Communicable Disease,” “Credit Card Number,” “Dependency or Neglect,” “Drug and Alcohol Treatment,” “Genetic Information,” “Juvenile Court History,” “Parole Status,” “Paternity,” “Student Grades or Performance,” “Tax Lien,” “Vehicle License Plate Number,” and “Voting Record.” Recall that nearly 40% of the documents contained an appendix.<sup>212</sup> Although there was substantial variability among the different

---

209. *See, e.g.*, Whiteman, *supra* note 17, at 470, 477 (describing the Northern Kentucky Law Library’s decision to refrain from scanning appendices to briefs filed in the Kentucky Supreme Court); Hinderman, *supra* note 18, at 6 (noting that the Montana State Law Library pulled electronic court records it had already posted online to remove all exhibits and appendices before reposting the briefs).

210. *See supra* note 184 and accompanying text.

211. The median frequency of sensitive information in briefs and appendices filed in criminal cases was 113 and 41, respectively; for juvenile cases, the median frequency of sensitive information was 17 and 6.5, respectively. The median frequency of sensitive information in briefs and appendices filed in civil cases was 16.

212. *See supra* Table 1.

document types with regard to the inclusion of an appendix,<sup>213</sup> the variation between case types with regard to appendices was less pronounced and the differences were not statistically significant.<sup>214</sup> Accordingly, we can conclude that it is not a dearth of documents with appendices in our sample that is suppressing the appearance of sensitive information in the appendices. We will return to the role of appendices in Part VI.

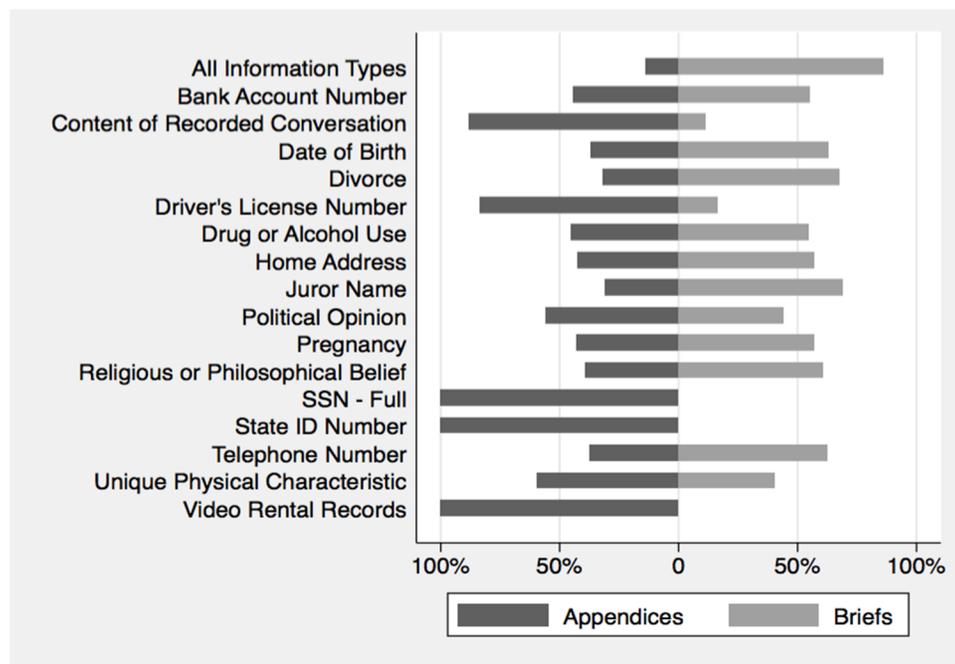


Figure 8: Horizontal bar graph showing percentage of sensitive information types in appendices and brief bodies. Only information types that appeared more than 30% of the time in appendices are listed.

### 3. Temporal Variations

Our final contextual factor is time. Following the approach of the North Carolina Supreme Court, which maintains its case files based on the year a case is decided by the court, we assigned the corresponding case year to each document in our sample. Figure 9 graphs the total number of

213. Nearly all petitions for discretionary review contained an appendix (98%) while briefs by the state were the least likely document type to include an appendix (16%). *See supra* Table 1.

214. Overall, 40% of civil cases included an appendix and 35% of criminal cases included an appendix. *See supra* Table 1. As noted in the text, the difference between the various case types with regard to the inclusion of appendices was not statistically significant (chi-square with 2 degrees of freedom = 2.1090,  $p = 0.348$ ).

documents in our sample by year, as well as the number of documents that did not contain any of the sensitive information types we coded for. As Figure 9 shows, there was considerable year-to-year variation in both of these measurements.

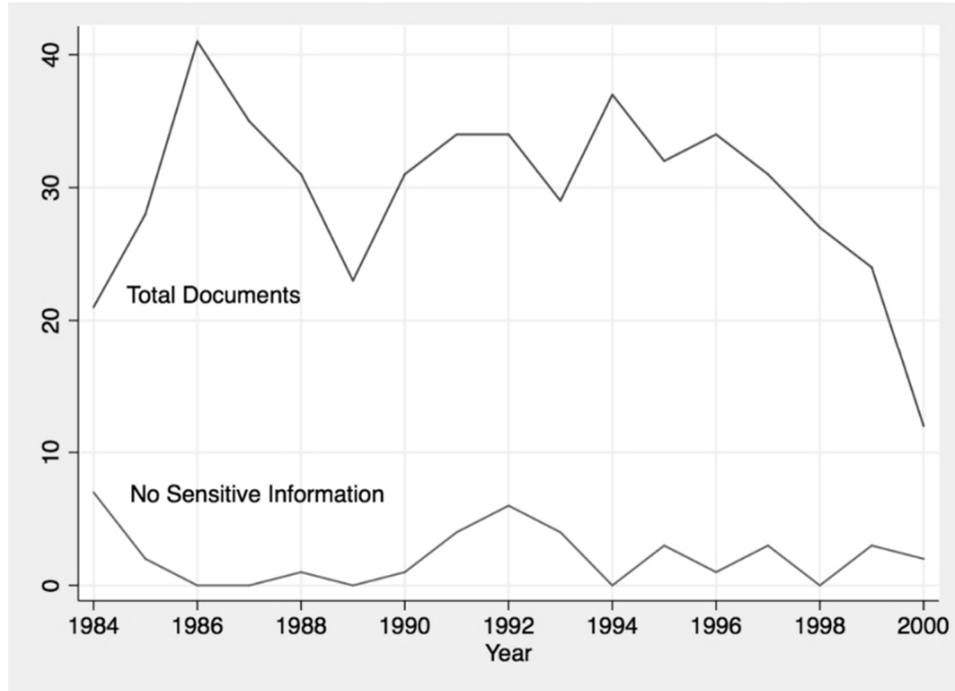


Figure 9: Total number of documents in sample and number of documents without sensitive information, by year.

We also found substantial variation in the total frequency of sensitive information per year. Figure 10 presents a two-way area graph of the total appearance of sensitive information by year as well as for the four most frequently occurring information categories: criminal proceedings, health, identity, and location. As Table 10 indicates, the total amount of sensitive information ranged from a low of 1,068 in 1984 to high of 6,052 in 1994, with an average of 3,117.5 pieces of sensitive information per year during the seventeen-year period under study.<sup>215</sup>

From Figure 10 we can see that in addition to variation in the overall frequency of sensitive information per year, the individual categories of sensitive information also varied during this time period. Not surprisingly,

215. The median frequency of sensitive information per year is 2,922; standard deviation is 1,700.3; and the interquartile range, covering the middle 50% of the observed frequencies, is 1,617–4,727.

information related to criminal proceedings tracked the overall totals quite closely (this is not surprising because the vast majority of sensitive information each year was associated with the criminal proceedings category). The other categories showed some variability as well, but did not parallel as closely the timing of the changes in the overall total. For example, the health category varied from a high of 626 in 1987 to a low of forty-four in 1997; the identity category varied from a high of 458 in 1995 to a low of fifty-seven in 1998; and the location category varied from a high of 305 in 1985 to a low of two in 1992.

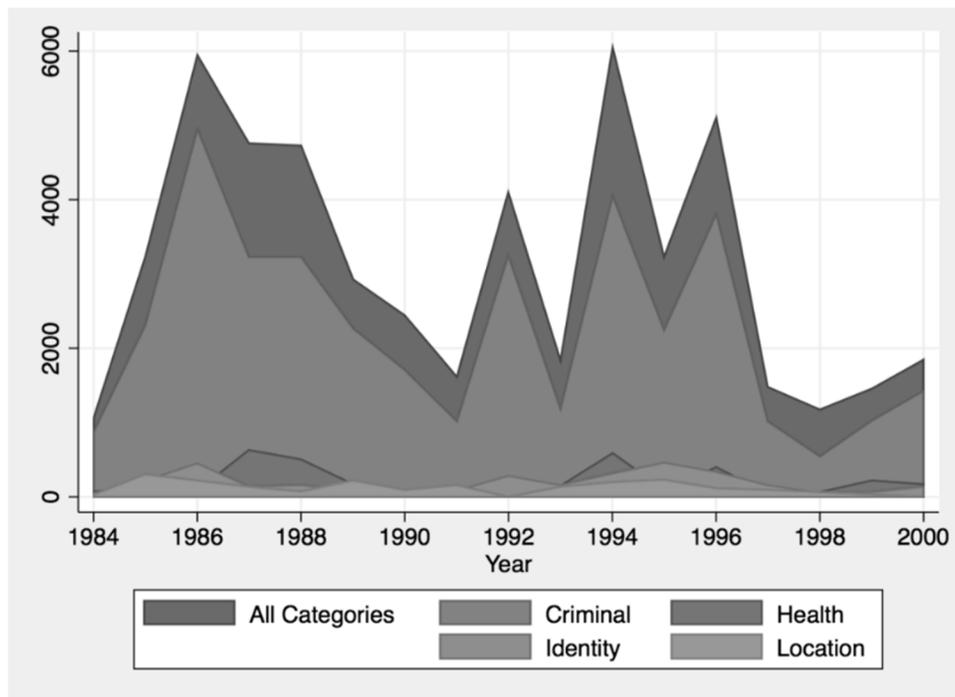


Figure 10: Total appearance of sensitive information by year of case decision including the 4 most frequently occurring information categories.

If we compare Figures 9 and 10, we might surmise that at least some of this variation is due to the fluctuations in the number of documents in our sample from each year. To remove this factor from our analysis, we calculated the frequency of appearance of the various categories on a per document basis. Figure 11 presents this measure of “sensitive information density” for the six most frequently occurring information categories.

Figure 11 suggests that there was no overarching trend in the appearance of sensitive information during the 1984 to 2000 time period. Instead, the numbers vary within a relatively constant range. It should be noted that the individual line graphs in Figure 11 do not all utilize the

same  $y$ -axis scale. The criminal proceedings category varied between twenty and 120 pieces of sensitive information per document whereas the location and civil proceedings categories varied between zero and ten. By varying the  $y$ -axis scale, Figure 11 allows us to compare the relative changes in sensitive information density *within* each category over time and leads to two important observations.

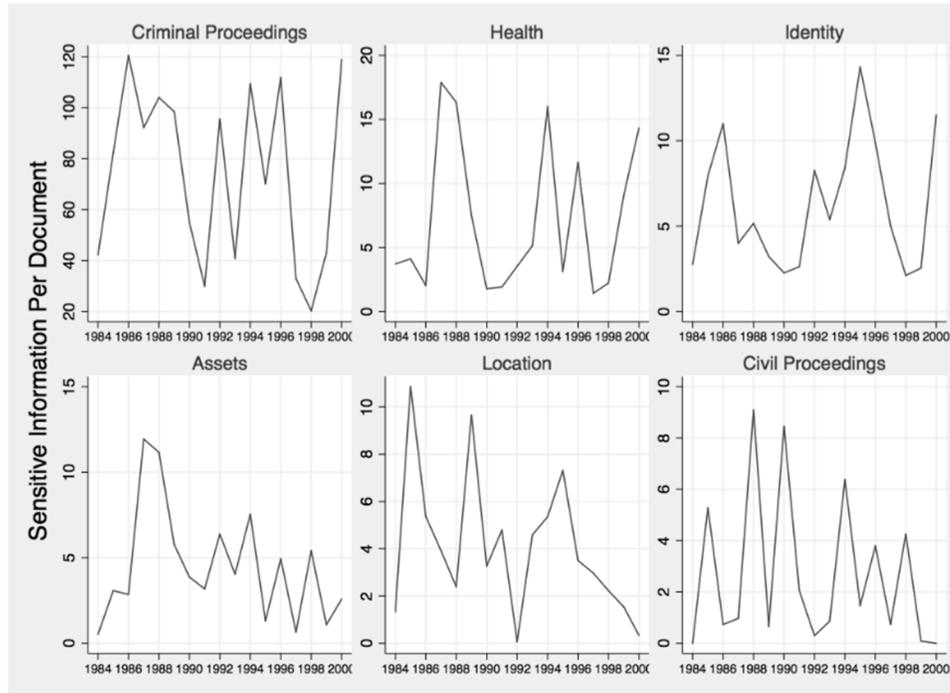


Figure 11: Sensitive information per document by year of case decision for the six most frequently occurring information categories.

First, there was considerable variability both within the categories and among the categories during this time period. None of the categories evidenced a consistent amount of sensitive information per document on a year-to-year basis. In some years there were sharp increases in the amount of sensitive information associated with these categories while in other years there were steep declines. For example, the amount of sensitive information per document in the health category spiked in 1987 and 1984 and fell in 1990 and 1997, whereas the identity category showed sharp increases in 1986 and 1995 and declines in 1990 and 1998. Moreover, the peaks and valleys evident in the individual graphs in Figure 11 do not align. When some categories were peaking, others were dipping.

Second, we do not see either a declining or rising trend for these categories. Although there is a pronounced increase in sensitive

information per document beginning in 1998 for the criminal proceedings, health, and identity categories, the location and civil proceedings categories declined during that same period, and overall the five-year moving averages for all of the categories show no discernable trend. From Figure 9 we can see that the upswing in sensitive information for the criminal, health, and identity categories in 1998 occurred during a period when the total number of documents was declining. Whether this increase in sensitive information continued after 2000 is beyond the scope of this study.

#### 4. Regression Analysis

In this section we use multiple regression analysis to examine whether and to what extent certain document and case characteristics influence the amount of sensitive information in court records. As we previously noted, document length is a statistically significant predictor of the amount of sensitive information in a court record.<sup>216</sup> Recall that the linear regression line shown in the scatterplot of sensitive information per document in Figure 4 provided an estimate of the relationship between the frequency of sensitive information in a document and the document's length.<sup>217</sup> We now add other independent variables to our analysis in order to better predict the amount of sensitive information in court records.

Figure 12 presents a nomogram of the multiple regression coefficients for the analysis of the frequency of sensitive information per document (log transformed) for nine independent case and document variables.<sup>218</sup> It shows that six independent variables—criminal case type, appellant's brief, appellee's brief, petition for discretionary review, state's brief, and document length (in pages)—are statistically significant predictors of the total amount of sensitive information in a document because their 95%

---

216. *See supra* note 195 and accompanying text.

217. The best fitting linear regression line in Figure 4 predicted a relationship of approximately 1:3 between document length and frequency of sensitive information, with a coefficient of determination,  $R^2$ , of 0.3299. *See supra* note 195 and accompanying text.

218. We utilized a log transformation of the total amount of sensitive information per document because this variable is not normally distributed. Each parameter estimate in Figure 12 is represented by a dot along with the 95% confidence interval for the estimate depicted by a horizontal line. Parameters with narrower confidence intervals are estimated more precisely than those with wider confidence intervals. Only those parameters with a confidence interval that does not cross zero are statistically significant at a 95% level. Figure 12 does not include the intercept parameter, which has a coefficient of 0.6542635 [95% CI: -1.089448 to 2.397975].

confidence intervals do not cross zero.<sup>219</sup> In other words, holding all other variables in the model constant, the amount of sensitive information in a document can be predicted based on the document's length, the type of brief (other than an amicus brief), and whether it was filed in a criminal case. The other variables listed in Figure 12 may also influence the amount of sensitive information in a document, but the findings from the multiple regression model do not show that we can be sufficiently confident to assess what effect, if any, they have on the amount of sensitive information.

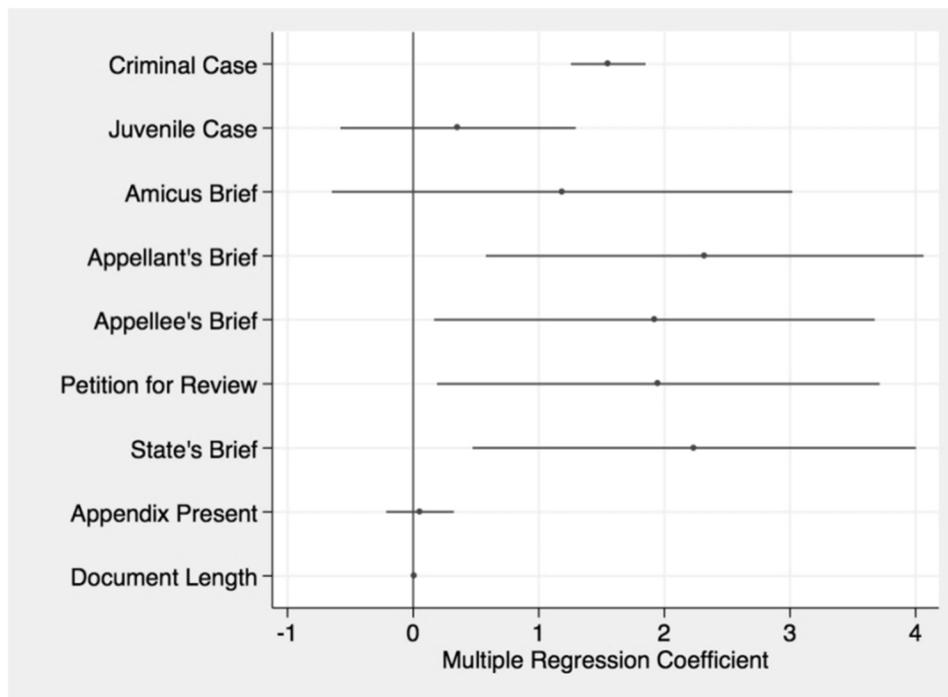


Figure 12: Nomogram of multiple regression parameters and 95% confidence intervals for the analysis of sensitive information per document (log transformed) listed by case and document variables.

What the multiple regression model tells us is that, other things being equal, criminal cases contain significantly more pieces of sensitive information per document than civil cases (coef. = 1.55,  $p = 0.000$ ). In addition, appellee briefs (coef. = 1.92,  $p = 0.032$ ), petitions for discretionary review (coef. = 1.95,  $p = 0.030$ ), briefs by the state (coef. =

219. The model's other relevant statistics include  $F(9,457) = 43.71$ ,  $p < 0.0000$ ,  $R^2 = 0.4626$ . All six independent variables described in the text added statistically significantly to the prediction,  $p < 0.05$ . The full regression table is included in the Appendix.

2.23,  $p = 0.013$ ), and appellant briefs (coef. = 2.32,  $p = 0.009$ ) contain significantly more pieces of sensitive information than the other briefs category (reference group), with the relative impact of the various brief types increasing as their coefficients increase. It also tells us that the number of pages in a document is another significant predictor of sensitive information. The more pages in a document, the more pieces of sensitive information appear in the document (coef. = 0.01,  $p = 0.000$ ). As for the inclusion of an appendix, the model shows that this does not predict the overall frequency of sensitive information in the document.

This final point about the lack of impact of appendices may strike some readers as surprising, but given our earlier finding that brief bodies contain a higher median frequency of sensitive information—as well as a wider variety of sensitive information types—than appendices,<sup>220</sup> it is not unexpected.<sup>221</sup> We will return to the implications of these findings in Part VI.

## VI. IMPLICATIONS FOR ACCESS POLICIES AND PRACTICES

As we noted in the introduction, courts and libraries are moving quickly to digitize court records and make them available online. The results of this study will, we hope, inform these efforts by providing much-needed detail about the extent and context of sensitive information in these important public records. In this part we discuss how our study can aid the ongoing debate about privacy and court records and how our results can help to identify and remedy potential implementation challenges if courts or archivists decide to carry out privacy management protocols.

It is not our goal in this Article to tell courts or archivists what information, if any, should be redacted or what documents should be withheld from online access or otherwise managed for privacy protection. These largely normative questions must be answered based on a careful balancing of the competing public access and privacy interests. Moreover, the privacy interests cannot be evaluated based solely on the presence or absence of specific types of sensitive information in individual court records. Other factors, including what one can learn or infer about individuals from other sources, as well as the value to society of the

---

220. See *supra* note 184 and accompanying text.

221. Another possible explanation is that the inclusion of an appendix is highly correlated with page length.

information in question, must also be taken into account. The data we present will be invaluable in doing this evaluation, but our findings should not be read to dictate one approach or another.

As discussed in Part IV, our sample of court documents came from a large corpus of briefs and other filings submitted to the North Carolina Supreme Court spanning the time period 1984 to 2000.<sup>222</sup> Although our results are specific to this population of court records, we believe that the data we collected shed light on court records held by other state appellate courts, particularly courts of last resort.<sup>223</sup>

Trial court records, however, are a different story. It is unlikely that our findings are generalizable to the records held by trial courts. Unlike appellate courts, which are primarily engaged in deciding questions of law, trial courts must resolve competing factual claims. As a result, their files likely contain a higher frequency of sensitive information from a wider array of records,<sup>224</sup> including pre-trial discovery materials, expert reports, juror questionnaires, court transcripts, physical evidence, and audiovisual materials.<sup>225</sup> Nevertheless, the methods we used in coding and analyzing the North Carolina Supreme Court's records could be applied to trial court records as well, and we hope that other researchers will do so.

---

222. See *supra* note 161 and accompanying text. We did not review the “record on appeal,” which is a separate filing containing, *inter alia*, copies of the case pleadings, jury instructions, transcripts, and other evidence entered in the lower courts.

223. Unfortunately, we could not find any data comparing state court appellate caseloads during the time period we studied. The National Center for State Courts (NCSC) did not implement the *State Court Guide to Statistical Reporting* until 2007, see *supra* note 180, and has not reclassified its historical data using the new reporting schema outlined in the guide. See Email to David Ardia from Shauna M. Strickland, Senior Court Research Analyst, NCSC, dated July 1, 2015 (on file with authors). Because of this, we cannot match our results with the data reported by other states during the time period we studied. We hope that other researchers will collect the data needed to do such comparisons.

224. We are aware of no comprehensive studies of trial court records. Carl Malamud, who used software to search for unredacted social security numbers in federal district court filings on PACER, noted that “often when our tool reported a Social Security number violation, when we looked around the document we also picked up many other Social Security numbers, birth dates, driver license numbers, Alien IDs, and bank account numbers.” Malamud Letter, *supra* note 6, at 1.

225. Some of these materials may end up in an appellate court's files if they are submitted as part of the record on appeal, but rarely do they appear to the same extent in the parties' briefs and appendices. See Conley et al., *supra* note 13, at 776 (noting that trial court records “contain an abundance of personal information, some of which may drop away as cases move from trial courts to appellate courts”).

### A. IDENTIFYING WHERE PRIVACY RISKS ARE GREATEST

We knew going into this study that court records contain sensitive information. Indeed, scholars have long argued that court records raise substantial privacy concerns. What we lacked, however, was comprehensive data about the extent and context of this information. These data are essential to understanding the threats to privacy that court records present.

In Part V, we identified the most common types of sensitive information that appear in the North Carolina Supreme Court's case files. In this section we begin to evaluate the potential "harmfulness" of this information, based on frequency of appearance, certain document and case attributes, and existing legal authority requiring or recommending redaction. We note at the outset that "harmfulness" is a contested concept in privacy law,<sup>226</sup> and we do not take a position in this Article as to how "harm" should be defined in the context of public records. Instead, we present an assessment of relative risk based on the frequency of occurrence and context of a broad range of sensitive information types. Regardless of how one defines the harm that comes from the disclosure of certain types of sensitive information, the harmfulness of that information will likely be influenced by the frequency of disclosure and its context both within court records and the larger information ecosystem.

The following sections highlight our key findings.

#### 1. *Court Records Vary Substantially in the Sensitive Information They Contain*

We found that court records vary substantially in both the types and frequency of sensitive information they contain. The records we studied did not exhibit every type of sensitive information in equal measure. Some information appeared much more often than other information. And some types of information that privacy advocates have highlighted did not appear at all in the records we studied. Moreover, nearly one in ten documents in our sample did not contain any of the 140 sensitive information types we coded for in this project.

---

226. As we noted in Part III, there is considerable disagreement among privacy scholars about the nature of the interests that privacy advances. In fact, some scholars argue that courts should abandon harm-based rationales entirely when evaluating privacy claims. See, e.g., James Peterson, *Behind the Curtain of Privacy: How Obscenity Law Inhibits the Expression of Ideas About Sex and Gender*, 1998 WIS. L. REV. 625, 632 n.38 (1998) ("In practice, the distinction between *harm* and *offense* is not always easy to maintain, because extreme forms of offense can cause emotional and even palpable physical harm.").

Although it is not surprising that certain types of sensitive information appeared more often than others, we were surprised by some of the patterns we found. To facilitate the collection and analysis of our data, we grouped the various information types into thirteen categories.<sup>227</sup> When we compared these categories, we found that sensitive information in seven categories appeared much more frequently than the other categories of information. The most commonly occurring categories—location, identity, criminal proceedings, health, assets, financial information, and civil proceedings—each appeared in at least 20% of the documents we studied; the remaining six categories each appeared in fewer than 8% of the documents.<sup>228</sup>

Information types in the less frequently appearing categories warrant comment because of their unexpected absence in the documents. Information about sexual practices appeared infrequently, as did information in the intellectual pursuits category, which includes religious beliefs, political opinions, and voting and reading records.<sup>229</sup> Information about education was also mostly absent from the documents as were photos and videos. None of the documents contained any sensitive information that fell within the computer use category (e.g., user names, passwords, and search history).

We also found substantial variability in how often certain types of sensitive information appeared in individual documents. For most of the information categories, sensitive information appeared between three and six times per document.<sup>230</sup> There were outliers, however. Information in the criminal proceedings category appeared far more frequently in the documents than any other information category, showing up approximately nine to eighteen times as often as the other categories.<sup>231</sup> Interestingly, although information in the images and sexual activities categories did not appear in very many documents, when it did appear, it appeared more often in a document than all the other categories except the criminal proceedings category.<sup>232</sup>

---

227. For a description of the categories we utilized, *see supra* Section III.B.

228. *See supra* Table 4 and accompanying text.

229. Intellectual pursuits are a category of information that only recently began to receive protection under U.S. privacy laws, but is the subject of an increasing amount of privacy scholarship. *See supra* Section III.B.11.

230. *See supra* Table 4 and accompanying text.

231. The “images” category is the other outlier, with a median of seventeen appearances per document. *See supra* Table 4.

232. *See supra* Section V.A.2 and accompanying text.

Again, it is just as important to focus on what we did not find in the records. The types of information that most people would associate with financial fraud and identity theft appeared less often than we expected. We found surprisingly few social security numbers, bank account numbers, credit card numbers, and other financial account numbers, each of which appeared in no more than three documents in our sample of 504 documents.<sup>233</sup> We also found no partial social security numbers, debit card numbers, credit reports, or personal identification numbers (PINs) in the records.

Several types of sensitive health information also appeared less frequently than we expected. Information about abortion, paternity, and communicable diseases each appeared in only one document. Genetic information appeared in only three documents, which may be due in part to the time period we studied (1984–2000),<sup>234</sup> and information about drug or alcohol treatment appeared in only four documents. We found no references in the documents to an identified individual having HIV or AIDS.

## 2. *Criminal Information Is Pervasive in Court Records*

What we did find in great numbers in the court records was information related to criminal proceedings, particularly witness names, crime victim names, arrests, criminal charges, and the names of subjects under investigation. Indeed, information in the criminal proceedings category pervaded the court records we reviewed. It not only appeared in most of the documents, it also appeared more often in those documents than any other category of sensitive information.

More than half of the documents we analyzed contained information that fell in the criminal proceedings category, and the individual information types in this category far outpaced every other type of information we coded for in terms of frequency of occurrence. The names of witnesses in criminal cases appeared more often than any other information type in our data set, followed by the name of an individual

---

233. *See supra* Figure 3 and accompanying text. Although the North Carolina Rules of Appellate Procedure do not require that social security numbers be excluded from briefs filed in the North Carolina Supreme Court, the rules do state that SSNs “shall be deleted or redacted from any document before including the document in the record on appeal.” N.C. R. APP. P. 9(a)(4). Our project did not involve the coding of the records on appeal.

234. *See* MICHAEL LYNCH ET AL., *TRUTH MACHINE: THE CONTENTIOUS HISTORY OF DNA FINGERPRINTING* 13 (2008) (noting that DNA evidence was first used in criminal investigations in the late 1980s but challenges in the courts and scientific press slowed its acceptance until the mid-1990s).

who was the subject of a criminal investigation. The name of a victim of criminal activity other than rape (rape victim name is a separate information type) was the third most frequently occurring information type, and information about arrests and convictions were the fourth and fifth most common information types, respectively.

Additionally, when information about criminal proceedings appeared in a document, it did so in large numbers. As we noted above, for most of the information categories sensitive information appeared between three and six times per document. By comparison, information in the criminal proceedings category appeared more than fifty times per document.<sup>235</sup> This substantially higher frequency of appearance is not due to a profusion of criminal cases. Documents from criminal cases made up only 36% of our sample.<sup>236</sup> Instead, what we see in the data is that criminal information appeared in documents filed in every type of case, including civil and juvenile cases.

Our data cannot tell us why criminal information is so pervasive in court records, but we can speculate based on qualitative factors. Information types in the criminal proceedings category—and in the civil proceedings category as well—relate to the functioning of the court system itself,<sup>237</sup> so it is perhaps not surprising to find these information types throughout the North Carolina Supreme Court's records. Indeed, when we examine the context of information from both of these categories, we see that information about the functioning of our criminal and civil courts is widely dispersed across all of the document types and case types we coded. As we noted, however, our data cannot definitely answer why criminal information is so common in court records. We hope that other researchers will take up this question.

### 3. *Criminal Cases Have Disproportionately More Sensitive Information*

Although documents from criminal cases constituted only slightly more than a third of the sample, they had an outsized impact on the types and frequency of sensitive information. More than three quarters (76.3%) of the sensitive information came from documents filed in criminal cases.

---

235. The criminal category appeared at a median frequency of appearance per document of 54.5. *See supra* Table 4 and accompanying text.

236. *See supra* Table 2.

237. Briefs filed in appellate courts often include arguments that turn on how the justice system functions, including acts and omissions by law enforcement, the credibility of witnesses, the relevancy of prior convictions and civil judgments, and the fairness of the jury.

And not only did criminal cases contain more sensitive information than civil and juvenile cases, they also contained a greater variety of sensitive information.

Criminal cases contained sensitive information from all of the categories that we identified in the documents,<sup>238</sup> whereas information from two categories, education and images, was entirely absent from civil cases.<sup>239</sup> Overall, criminal cases contained more individual types of sensitive information than civil and juvenile cases. Of the 140 sensitive information types that we coded for in the records, ninety-five distinct types appeared in the documents. Although some types appeared exclusively or nearly so in civil cases, documents filed in criminal cases contained a greater number of sensitive information types.<sup>240</sup>

Criminal cases also had substantially more sensitive information per document than either civil or juvenile cases. In fact, the median frequency of sensitive information in documents filed in criminal cases was approximately five times greater than that of documents filed in either civil or juvenile cases.<sup>241</sup> As a result, criminal cases had a higher density of sensitive information per page than either civil or juvenile cases. As page length increased, the number of pieces of sensitive information in criminal cases increased at a higher rate than it did in civil and juvenile cases. For criminal cases, the ratio between page length and frequency of sensitive information was four times greater than that of civil and juvenile cases.

One of the drivers of this disparity may be the disproportionate impact of death penalty cases.<sup>242</sup> Although documents from death penalty cases constituted only 6.5% of our sample,<sup>243</sup> they contained more than a quarter

---

238. We did not find any information in the documents that fell within the “computer use” category. *See supra* Table 4.

239. Six of the thirteen information categories were absent from documents filed in juvenile cases.

240. *See supra* Table 5 and accompanying text.

241. Sensitive information in criminal cases also appeared much more frequently in the brief body than in the appendix. *See supra* Figure 2.

242. During the period from 1989–1998, there was a sharp increase in the number of documents from cases in which a defendant was sentenced to death (from 0.8 documents per year to 2.7 per year). This corresponded to an increase in death penalty verdicts in North Carolina, which rose from nine in 1989 to thirty-four in 1995. CTR. FOR DEATH PENALTY LITIG., ON TRIAL FOR THEIR LIVES: THE HIDDEN COSTS OF WRONGFUL PROSECUTIONS IN NORTH CAROLINA 16 (2015), <http://www.cdpl.org/wp-content/uploads/2015/06/INTERACTIVE-CDPL-REPORT.pdf>. In North Carolina, death penalty convictions are subject to automatic review by the North Carolina Supreme Court. N.C. GEN. STAT. ANN. § 15A-2000(d)(1) (West 2015).

243. The most commonly occurring appellate subject area was “tort, contract, and real property,” which constituted 32% of the documents in the sample, followed by

(27.7%) of the total amount of sensitive information. Documents from death penalty cases also were, on average, the longest documents in the sample.<sup>244</sup> Interestingly, most of the appearances of sensitive information types in the education category occurred in documents filed in death penalty cases, whereas information types in the employment and financial categories were largely absent from documents filed in these cases.<sup>245</sup>

#### 4. *Minors Deserve Additional Attention*

The vast majority of the sensitive information was associated with adults. Nevertheless, because of heightened concerns about the privacy of children, information associated with minors deserves special attention.<sup>246</sup>

As the data showed, sensitive information about minors was not limited to juvenile cases. Overall, 7% of the sensitive information was associated with an identified minor. Criminal cases evidenced slightly more information associated with minors than civil cases, but the difference was not substantial.<sup>247</sup> Not surprisingly, juvenile cases contained significantly more information about minors: 40% of the sensitive information in juvenile cases was associated with a minor.<sup>248</sup>

---

“felony (non-death penalty),” which arose in 26% of the documents. *See supra* notes 180–181 and accompanying text.

244. Documents filed in death penalty cases had a mean length of 113.63 pages, more than twice the average length for all documents in the sample (sample mean was 47.92 pages).

245. The higher frequency of appearance of information relating to education in death penalty cases may be due to its relevance as mitigation evidence. *See* N.C. GEN. STAT. § 15A-1340.16(e). Documents in death penalty cases also had the highest incidences of the following sensitive information types: “Criminal Sentence,” “Incarceration,” “Juror Names,” and “Student Discipline.”

246. Some commentators have argued that all juvenile proceedings should be closed by default in order to protect the interests of minors. *See, e.g.,* William Wesley Patton & Kelly Crecco, *An Update to Striking a Balance: Freedom of the Press Versus Children’s Privacy Interests in Juvenile Dependency Proceedings*, 12 FIRST AMEND. L. REV. 575, 589 (2014) (“Children are at much more risk of juragenic psychological harm in a presumptively open dependency court system than in a discretionarily open court system for a number of reasons.”); *see also* Kristin Henning, *What’s Wrong with Victims’ Rights in Juvenile Court?: Retributive Versus Rehabilitative Systems of Justice*, 97 CALIF. L. REV. 1107, 1158–60 (2009) (suggesting there can be tensions between victim rights and the confidentiality of juvenile prosecutions).

247. Sensitive information in criminal cases was associated with minors 7.2% of the time. In civil cases it was 6.1%.

248. *See supra* Section V.B.2.b. Although we did not record whether the minor in question was a “covered juvenile” under North Carolina law and thus entitled to additional privacy protections, a recent study published by the Juvenile Law Center found that the vast majority of states—including North Carolina—are failing to protect highly sensitive information contained in juvenile court records. JUVENILE LAW CTR., FAILED

Although the overall amount of information associated with minors was low, there were some notable differences between the information categories with regard to minors. Relative to their baseline percentage across all categories, minors were less likely to be associated with information relating to criminal proceedings and more likely to be associated with information in the education, health, identity, and sexual activities categories.<sup>249</sup> In addition, a number of specific information types were disproportionately associated with minors. Information about “Communicable Disease” and, unsurprisingly, “Name of Minor Child,” for example, were exclusively identified with minors,<sup>250</sup> and several information types associated with both adults and minors appeared more often with minors than the overall percentages would have suggested, including: “Adoption,” “Custody or Guardianship,” “Date of Birth,” “Juvenile Court History,” “Pregnancy,” and “Sex Life.”

On the other hand, there were several information types we expected to find associated with minors more often than we observed. For instance, no photographs or videos were associated with minors. Information about “Student Discipline” appeared in only one document, and “Student Grades and Performance Evaluations” appeared in only six documents.

Given the small number of documents from juvenile cases in the sample ( $n = 7$ ), the inferences regarding the extent of sensitive information associated with minors is limited, especially in juvenile cases. A larger sample of documents from juvenile cases is necessary to better understand the privacy risks associated with minors. This is also an area we hope future researchers will explore.

##### 5. *It Is Unwise to Focus Exclusively on Appendices*

When we began this study we assumed, based largely on anecdotal reports from archivists and privacy scholars, that appendices included in court records would contain more sensitive information than legal briefs and that highly sensitive information that had been kept out of legal briefs would nevertheless appear in the appendices. Our study showed, however, that appendices are for the most part not quantitatively different from

---

POLICIES, FORFEITED FUTURES: A NATIONWIDE SCORECARD ON JUVENILE RECORDS (2015), <http://jlc.org/blog/new-study-reveals-majority-us-states-fail-protect-juvenile-records>.

249. See *supra* Section V.B.2.b.

250. Information concerning “Communicable Disease” appeared only once in the sample. As noted in Part III, we coded for “Name of Minor Child” as a specific information type; this was the most common information type associated with minors. See *supra* Section V.B.2.b, especially Figure 7.

legal briefs in terms of the frequency and types of sensitive information they contain.

In terms of the amount of sensitive information in a document, the data actually showed that legal briefs contained a higher frequency of sensitive information than appendices. This disparity was particularly evident in criminal and juvenile cases, where the brief bodies contained approximately three times as much sensitive information as the appendices; for civil cases, sensitive information appeared with equal frequency in the appendices and briefs.<sup>251</sup> These findings were reinforced by the multiple regression model, which showed that a document's inclusion of an appendix did not affect the total amount of sensitive information in the document.<sup>252</sup>

With regard to the specific types of sensitive information in briefs and appendices, the results were more mixed. Only seven of the ninety-five information types that we identified in the records appeared more often in the appendices.<sup>253</sup> On the other hand, twenty-seven information types appeared exclusively in the briefs, and many more appeared more than 50% of the time in the briefs. Nevertheless, the information types that did appear more often in the appendices were the types of information many would regard as particularly sensitive from the standpoint of identity theft. Uniquely identifying physical characteristics, drivers' license numbers, social security numbers, and state identification numbers all appeared more often in the appendices, with the latter three information types appearing exclusively in an appendix.

Accordingly, although there is good reason to pay careful attention to appendices when reviewing court records for sensitive information, it is unwise to focus exclusively on appendices. More types of sensitive information appear in legal briefs, and at a higher overall frequency than in appendices.

#### 6. *Trends in Sensitive Information over Time*

Although the amount of sensitive information in the case files of the North Carolina Supreme Court varied significantly during the time period of this study (1984–2000), there were no overarching trends in the frequency of sensitive information during this seventeen-year period.<sup>254</sup>

---

251. *See supra* Figure 2 and accompanying text.

252. *See supra* Section V.B.4.

253. *See supra* Figure 8 and accompanying text.

254. *See supra* Figure 10 and accompanying text.

Instead, the most commonly occurring information types appeared with a frequency that varied within a relatively consistent range.

This is not to say that the amount of sensitive information was constant during this time period. To the contrary, there was a great deal of year-to-year variability. In some years there were sharp increases in the amount of sensitive information while in other years there were steep declines. Moreover, the various categories of information did not rise and fall together.

Nevertheless, these variations, which appear to be cyclical, do not show a declining or rising trend during the time period under study. Whether this would continue to be the case if we extended our collection of court records earlier or later in time is beyond the scope of this study.

A number of important events occurred in the late 1990s and early 2000s that might have a significant impact on the frequency and extent of sensitive information in court records in the years following our study. In 1999, the North Carolina court system began allowing electronic filing and in 2009 implement e-filing rules that placed the onus on the parties to redact a number of types of sensitive information from case filings, including social security numbers and certain financial information.<sup>255</sup> In addition, in the late 1990s there was a significant rise in the use of computers and electronic communication systems that might have led to the generation of different types of sensitive information in court records. Given the time it takes for a case to work its way up to a state's highest court, we can expect that these changes would likely take a few years to be reflected in the North Carolina Supreme Court's files.

#### B. CHALLENGES IN IMPLEMENTING PRIVACY PROTECTIVE PRACTICES

In addition to aiding our understanding of privacy in the context of court records, the results of this research and the experience of the coders will have practical implications for court personnel and archivists as they develop rules and practices for electronic filing of court records or the digitization of older records. Although all of our findings should have some bearing on these efforts, we highlight four main points.

First, some types of sensitive information are easier to identify in court records than others. If courts or archivists decide to limit access to certain

---

255. See generally Deborah Leonard Parker, *Electronic Filing in North Carolina*, 2 J. APP. PRAC. & PROCESS 351 (2000) (describing the impact of the introduction of electronic filing rules in North Carolina in 1999). See also *supra* note 10 and accompanying text.

types of sensitive information, they will inevitably have to make choices about which information types should be restricted based, at least in part, on whether the burden of addressing privacy is commensurate with the risk of harm and whether investments in privacy protection practices will be effective. Such proportionality considerations get to the heart of the debate about the nature of privacy harms and the risks, both foreseeable and perhaps unforeseeable, that are presented by current and future practices of information aggregation and use. Without solving the normative and broader practical problems, though, this study nonetheless reveals that some of the sensitive information types in court records appear in standard formats—such as social security numbers, dates of birth, and financial account numbers—and therefore are more easily identifiable through automated searching techniques.<sup>256</sup>

Second, many sensitive information types require human readers to review records more than once in order to identify the information. Our coders reported that some sensitive information was identifiable only by reading the record and developing a sense of the narrative. For example, the first mention of an individual might not reveal that she was indeed a cooperating defendant or that a particular named person was a minor. This type of information would likely require a human reader to review the record and make note of names associated with sensitive information. Coders also reported that reading a document once would not necessarily be sufficient to capture all occurrences of names that could be associated with sensitive information. Even after this investment of human effort, the question remains about how best to balance privacy and judicial transparency if the sensitive information arises through the course of the narrative. The name itself might be amenable to redaction, but the story of cooperation might be too interwoven in a brief or appendix to make redaction feasible.

Third, redaction may be a poor strategy for dealing with some sensitive information types. Our coders reported that some court briefs and

---

256. See Rebecca Green, *Petitions, Privacy, and Political Obscurity*, 85 TEMP. L. REV. 367, 406 n.272 (2013) (noting that courts and judicial administrators have explored redaction using computer software, but also noting the prevalence of redaction errors in the federal court records PACER database); Ronald Leighton, Joe Cecil, Michael Ishakian & Edward Felten, *Panel Three: Implementation—What Methods, If Any, Can be Employed to Promote the Existing Rules' Attempts to Protect Private Identifier Information from Internet Access?* 79 FORDHAM L. REV. 45, 49 (2010) (Felten discusses the amenability of social security numbers to software redaction because of their fixed pattern and suggests that advanced machine learning methods could be developed to help locate and redact even “difficult types of information, such as names of minor children.”).

accompanying appendices revealed information that was difficult to code as a discrete occurrence. The story of child abuse, for example, might pervade a particular case record. Options for addressing privacy would need to account for this challenge. The preliminary coding also revealed that the occurrence of language conveying a person's gender is a poor fit for redaction because gender is so integral to the English language. As explained in Part III, the coding for gender threatened to overwhelm the coding process, and the same would be true for any redaction effort.

Fourth, the data showed that there might be some value in prioritizing the review of certain documents when searching for sensitive information in appellate court records. We found that the various brief types were not all equal in the amount of sensitive information they contained. For example, briefs filed by the state had the highest frequency of sensitive information, whereas amicus briefs had the fewest appearances of sensitive information.<sup>257</sup> In addition, documents filed in death penalty cases had a disproportionately higher rate of sensitive information than other types of cases. For court personnel and archivists seeking to make the best use of limited resources, it may make sense to focus on some types of documents and cases over others. We would caution, however, against focusing exclusively on appendices.<sup>258</sup>

## VII. CONCLUSION

Court records present a special challenge for privacy advocates. Unlike in many other areas of privacy law, the information in court records is presumptively open to the public. This openness serves many salutary functions, such as ensuring that our system of justice functions fairly and is accountable to the public. The public's right of access to court records and the information they contain, however, is not absolute.

Courts can—and frequently do—restrict public access when an overriding interest supports closure or sealing of specific information. Although the precise standard that a court must apply will vary depending on the circumstances, in general courts must conclude that the interest in prohibiting disclosure outweighs the strong presumption of public access. In the context of libraries and other archives, which may not be bound by

---

257. Briefs filed by the state had a median of 105.5 appearances of sensitive information per document; for amicus briefs, the median was 10.5. *See supra* Table 3. Multiple regression modeling also showed that the brief types (other than amicus briefs) were statistically significant predictors of the amount of sensitive information in a document.

258. *See supra* Section VI.A.5.

law to provide public access to court records, the question is not about what the law requires but about what policy best ensures the protection of privacy interests while simultaneously informing the public about the functioning of the court system.

Although we found a substantial amount of sensitive information in the court records we studied, we have not sought to tell courts or archivists what information, if any, should be redacted or what documents should be withheld from online access or otherwise managed for privacy protection. These largely normative questions must be answered based on a careful balancing of the competing public access and privacy interests. The data presented in this study will be helpful in this balancing, but the findings should not be read to dictate one approach or another.

Privacy interests cannot be evaluated based solely on the presence or absence of specific types of sensitive information in a single court document. Other factors, including the context of the information and the extent of information about an individual that is available from other sources, must also be taken into account. On that latter point, this study is but one piece in a complicated mosaic.

What this study has shown is that court records vary significantly in the types of sensitive information they contain. Records in civil cases are not identical to records in criminal cases or juvenile cases. Consequently, when scholars and policymakers discuss privacy and court records, they must be cautious of generalizing. Depending on the privacy concerns considered paramount, we are likely to see a very different risk profile between different types of court records, cases, and levels of the court system.

Much work remains in order to understand the privacy risks that might arise from online access to court records. We hope that future researchers will answer some of the questions that the data have raised, including the prevalence of criminal information in court records, the differences between appellate court records and trial court records, the extent of sensitive information in juvenile court files, and the impact of e-filing procedures on the types and frequency of sensitive information in court records.

## APPENDIX

Table A1: Sensitive information types in coding list, with category and number of documents that contained each information type, percentage of the total sample of documents ( $n = 504$ ), overall frequency of appearance, and percentage of total frequency of all sensitive information identified in the sample.

Sensitive Information Type	Category	Documents		Frequency	
		<i>n</i>	%	<i>n</i>	%
Abortion	Health	1	0.2%	1	0.0%
Adoption	Civil Proceedings	4	0.8%	114	0.2%
Age	Identity	84	16.7%	350	0.7%
Arrest or Charge	Criminal Proceedings	160	31.7%	2,512	4.7%
Asset-Other	Assets	3	0.6%	6	0.0%
Bank Account Number	Financial	3	0.6%	9	0.0%
Bankruptcy	Financial	4	0.8%	10	0.0%
Cable Television Subscription Record	Intellectual Pursuits	1	0.2%	4	0.0%
Cable Television Viewing History	Intellectual Pursuits	0	-	0	-
Cause of Death	Health	70	13.9%	174	0.3%
Child Abuse	Criminal Proceedings	27	5.4%	378	0.7%
Child Support	Civil Proceedings	11	2.2%	40	0.1%
Civil Commitment	Civil Proceedings	3	0.6%	27	0.1%
Civil Proceedings-Other	Civil Proceedings	2	0.4%	5	0.0%
Communicable Disease	Health	1	0.2%	1	0.0%
Compensation	Financial	125	24.8%	909	1.7%
Computer Use - Other	Computer Use	0	-	0	-
Content of Recorded Conversations	Intellectual Pursuits	3	0.6%	119	0.2%
Conviction	Criminal Proceedings	150	29.8%	1,744	3.3%
Cooperating Defendant Name	Criminal Proceedings	32	6.3%	772	1.5%
Credit Card Number	Financial	1	0.2%	2	0.0%
Credit Report	Financial	0	-	0	-
Criminal Proceedings-Other	Criminal Proceedings	17	3.4%	257	0.5%
Custody or Guardianship	Civil Proceedings	17	3.4%	257	0.5%
Date of Birth	Identity	30	6.0%	46	0.1%
Date of Death	Identity	85	16.9%	325	0.6%
Date of Hospital Stay	Health	37	7.3%	115	0.2%
Debit Card Number	Financial	0	-	0	-
Debt	Financial	7	1.4%	47	0.1%
Dependency or Neglect	Civil Proceedings	2	0.4%	15	0.0%
Disability Status	Health	24	4.8%	136	0.3%

Sensitive Information Type	Category	Documents		Frequency	
		<i>n</i>	%	<i>n</i>	%
Discipline	Employment	20	4.0%	237	0.5%
Divorce	Civil Proceedings	35	6.9%	200	0.4%
Domestic Violence Victim Name	Criminal Proceedings	7	1.4%	525	1.0%
Driver's License Number	Identity	2	0.4%	6	0.0%
Drug or Alcohol Dependency	Health	34	6.7%	325	0.6%
Drug or Alcohol Treatment	Health	4	0.8%	6	0.0%
Drug or Alcohol Use	Health	8	1.6%	135	0.3%
Education-Other	Education	1	0.2%	4	0.0%
Eligibility for School Lunch Program	Education	0	-	0	-
Email Address	Identity	0	-	0	-
Employment-Other	Employment	0	-	0	-
Fax Number	Identity	0	-	0	-
Financial Aid Award	Education	0	-	0	-
Financial Asset	Assets	74	14.7%	492	0.9%
Financial-Other	Financial	5	1.0%	11	0.0%
Fingerprint	Identity	0	-	0	-
Foreclosure Judgment	Financial	6	1.2%	81	0.2%
Full-Face Photograph	Images	1	0.2%	2	0.0%
Gait	Identity	0	-	0	-
Gender Identity Change	Identity	0	-	0	-
Genetic Information	Health	3	0.6%	8	0.0%
Geolocation Information	Location	52	10.3%	301	0.6%
Gun Permit	Assets	0	-	0	-
Gun Permit Application	Assets	0	-	0	-
Gun Possession or Ownership	Assets	65	12.9%	781	1.5%
Health Plan Beneficiary Number	Health	0	-	0	-
Health-Other	Health	65	12.9%	781	1.5%
HIV / AIDS Status	Health	0	-	0	-
Home Address	Location	55	10.9%	216	0.4%
Identity-Other	Identity	13	2.6%	32	0.1%
IM ID	Computer Use	0	-	0	-
Images-Other	Images	0	-	0	-
Incarceration	Criminal Proceedings	49	9.7%	295	0.6%
Informant Name	Criminal Proceedings	4	0.8%	290	0.6%
Insurance Policy Number	Financial	5	1.0%	14	0.0%

Sensitive Information Type	Category	Documents		Frequency	
		<i>n</i>	%	<i>n</i>	%
Intellectual Pursuits-Other	Intellectual Pursuits	1	0.2%	1	0.0%
Internet Protocol (IP) Address	Computer Use	0	-	0	-
Internet Search History	Computer Use	0	-	0	-
ISP Records	Computer Use	0	-	0	-
Iris Print	Identity	0	-	0	-
Juror Name	Criminal/Civil Proceedings	35	6.9%	1,563	3.0%
Juvenile Court History	Criminal Proceedings	4	0.8%	19	0.0%
Loan Account Number	Financial	0	-	0	-
Location-Other	Location	101	20.0%	1,122	2.1%
Medical Billing Number	Health	0	-	0	-
Medical Condition	Health	101	20.0%	1,122	2.1%
Medical Device ID/Serial Number	Health	0	-	0	-
Medical Record Number	Health	0	-	0	-
Military Discharge	Employment	8	1.6%	11	0.0%
Mother's Maiden Name	Identity	5	1.0%	9	0.0%
Mug Shot	Criminal Proceedings	0	-	0	-
Name of Minor Child	Identity	85	16.9%	1,581	3.0%
Name of Subject of Investigation	Criminal Proceedings	173	34.3%	8,284	15.6%
Other Crime Victim Name	Criminal Proceedings	136	27.0%	4,988	9.4%
Other Financial Account Number	Financial	2	0.4%	4	0.0%
Other Health Diagnosis or Treatment	Health	99	19.6%	1,397	2.6%
Parole Status	Criminal Proceedings	5	1.0%	24	0.1%
Passport Number	Identity	0	-	0	-
Password	Computer Use	0	-	0	-
Paternity Test	Health	1	0.2%	1	0.0%
Performance Evaluation	Employment	16	3.2%	59	0.1%
Personal Identification					
Code/Password	Financial	0	-	0	-
Photos/Videos - Fully Undressed	Images	0	-	0	-
Photos/Videos - Partially Undressed	Images	0	-	0	-
Photos/Videos -Violence, Abuse, Death	Images	1	0.2%	15	0.0%
Place of Birth	Identity	6	1.2%	7	0.0%
Place of Death	Health	28	5.6%	41	0.1%
Political Opinion	Intellectual Pursuits	2	0.4%	9	0.0%
Pregnancy	Health	14	2.8%	56	0.1%

Sensitive Information Type	Category	Documents		Frequency	
		<i>n</i>	%	<i>n</i>	%
Prescriptions	Health	7	1.4%	17	0.0%
Presentence Investigation Report	Criminal Proceedings	1	0.2%	1	0.0%
Prior Civil Judgment	Civil Proceedings	60	11.9%	528	1.0%
Professional Cert. or License Number	Identity	29	5.8%	76	0.1%
Racial or Ethnic Origin	Identity	19	3.8%	175	0.3%
Rape Victim Name	Criminal Proceedings	43	8.5%	1,121	2.1%
Real Estate Ownership/Rental	Assets	82	16.3%	1,103	2.1%
Records of Library Use	Intellectual Pursuits	0	-	0	-
Records of Reading Material	Intellectual Pursuits	3	0.6%	5	0.0%
Religious or Philosophical Belief	Intellectual Pursuits	16	3.2%	172	0.3%
RFID	Computer Use	0	-	0	-
School Address	Location	0	-	0	-
Sentence	Criminal Proceedings	3	0.6%	5	0.0%
Sex Life	Sexual Activities	16	3.2%	172	0.3%
Sex Video	Sexual Activities	0	-	0	-
Sexual Abuse Allegation	Criminal Proceedings	3	0.6%	12	0.0%
Sexual Activities-Other	Sexual Activities	129	25.6%	776	1.5%
Signature	Identity	0	-	0	-
SSN - Full	Identity	1	0.2%	1	0.0%
SSN - Partial	Identity	0	-	0	-
State ID Number	Identity	1	0.2%	1	0.0%
Student Discipline	Education	1	0.2%	1	0.0%
Student Grades or Performance Evals.	Education	6	1.2%	16	0.0%
Student ID	Education	0	-	0	-
Tax Lien	Financial	5	1.0%	5	0.0%
Tax Return	Financial	1	0.2%	5	0.0%
Telephone Number	Identity	231	45.8%	570	1.1%
Trade Union Membership	Intellectual Pursuits	0	-	0	-
Unique Physical Characteristic	Identity	5	1.0%	27	0.1%
User Name	Computer Use	0	-	0	-
Vehicle Identification Number	Assets	0	-	0	-
Vehicle License Plate Number	Assets	2	0.4%	3	0.0%
Victim of Identity Theft	Financial	0	-	0	-
Video Rental Records	Intellectual Pursuits	1	0.2%	1	0.0%
Voice Print	Identity	0	-	0	-

Sensitive Information Type	Category	Documents		Frequency	
		<i>n</i>	%	<i>n</i>	%
VOIP ID	Computer Use	0	-	0	-
Voting Record	Intellectual Pursuits	1	0.2%	2	0.0%
Witness Name	Criminal/Civil Proceedings	221	43.8%	14,437	27.2%
Work Address	Location	253	50.2%	752	1.4%
Zip Code	Location	202	40.1%	853	1.6%

Table A2: Regression results for the analysis of sensitive information per document (log transformed).

Variable	Coefficient	Standard Error
Criminal Case Type	1.552622*	0.1512302
Juvenile Case Type	0.3569533	0.4766879
Amicus Curiae Brief	1.185187	0.9326415
Appellant's Brief	2.32089*	0.8866871
Appellee's Brief	1.920483*	0.8927833
Petition for Discretionary Review	1.951337*	0.8965791
Brief for the State	2.237307*	0.8981081
Appendix Present	0.0542404	0.1371782
Document Length (in pages)	0.0109019*	0.0011283
Constant	0.2973103	0.8927754
<i>R</i> -squared	0.4626	
No. of observations	467	

\* indicates statistical significance ( $P < 0.05$ )



# PUSH, PULL, AND SPILL: A TRANSDISCIPLINARY CASE STUDY IN MUNICIPAL OPEN GOVERNMENT

*Jan Whittington, Ryan Calo, Mike Simon, Jesse Woo,  
Meg Young & Peter Schmiedeskamp<sup>†</sup>*

## ABSTRACT

Municipal open data raises hopes and concerns. The activities of cities produce a wide array of data, data that is vastly enriched by ubiquitous computing. Municipal data is opened as it is pushed to, pulled by, and spilled to the public through online portals, requests for public records, and releases by cities and their vendors, contractors, and partners. By opening data, cities hope to raise public trust and prompt innovation. Municipal data, however, is often about the people who live, work, and travel in the city. By opening data, cities raise concern for privacy and social justice.

This article presents the results of a broad empirical exploration of municipal data release in the City of Seattle. In this research, parties affected by municipal practices expressed their hopes and concerns for open data. City personnel from eight prominent

---

DOI: <http://dx.doi.org/10.15779/Z38PZ61>

© 2015 Jan Whittington, Ryan Calo, Mike Simon, Jesse Woo, Meg Young & Peter Schmiedeskamp.

<sup>†</sup> Jan Whittington is an Associate Professor of Urban Design and Planning in the College of Built Environments, University of Washington, and the Director of the Urban Infrastructure Lab. Ryan Calo is an Assistant Professor in the University of Washington School of Law, and Director of the University of Washington Tech Policy Lab. Mike Simon is Chief Technology Officer for Creation Logic, LLC. Jesse Woo is a Corporate Attorney in Berkeley, California and a Consultant at the University of Washington Tech Policy Lab. Meg Young is a Ph.D. Student in the Information School, University of Washington. Peter Schmiedeskamp is an Interdisciplinary Ph.D. Student in Planning at the University of Washington.

This project was conducted in partnership with the City of Seattle. The authors acknowledge Michael Mattmiller, Ryan Biava, Ginger Armbruster, Bruce Blood, and the many additional employees, residents, and business representatives of the City of Seattle, who generously gave their time to participate in this research. For their comments on this study, the authors would also like to thank the participants of the 19th Annual Berkeley Center for Law & Technology and Berkeley Technology Law Journal Symposium, Open Data: Addressing Privacy, Security, and Civil Rights Challenges, held on April 17, 2015 and Responsible Use of Open Data: Government and the Private Sector, held at New York University on November 19–20, 2015, co-organized by BCLT and NYU's Information Law Institute and Department of Media, Culture and Communication. This project was one of six funded, in part, by Berkeley Center for Law & Technology with a generous grant from Microsoft, with funding also provided by the City of Seattle.

departments described the reasoning, procedures, and controversies that have accompanied their release of data. All of the existing data from the online portal for the city were joined to assess the risk to privacy inherent in open data. Contracts with third parties involving sensitive or confidential data about residents of the city were examined for safeguards against the unauthorized release of data.

Results suggest the need for more comprehensive measures to manage the risk latent in opening city data. Cities should maintain inventories of data assets, produce data management plans pertaining to the activities of departments, and develop governance structures to deal with issues as they arise—centrally and amongst the various departments—with ex ante and ex post protocols to govern the push, pull, and spill of data. In addition, cities should consider conditioned access to pushed data, conduct audits and training around public records requests, and develop standardized model contracts to protect against the spill of data by third parties.

## TABLE OF CONTENTS

I.	INTRODUCTION.....	1902
A.	THE MUNICIPALITY IN FOCUS .....	1903
B.	PURPOSE, THEMES, AND CONTENT .....	1904
II.	OUR APPROACH .....	1905
III.	FINDINGS .....	1907
A.	QUALITATIVE ASSESSMENT I: KEY STAKEHOLDERS .....	1908
1.	<i>Methods: Data Collection and Analysis</i> .....	1909
a)	Research Design and Sampling .....	1909
b)	Data Collection .....	1910
c)	Data Analysis.....	1911
2.	<i>Findings</i> .....	1912
a)	Effects of Open Data Initiative on Public Trust .....	1912
b)	Economic Value Latent in Data.....	1912
c)	City Management of Open Data Initiative.....	1913
d)	Privacy Interests in Open Data .....	1914
e)	Safety Risks Latent in Data.....	1916
f)	Lack of Public Trust in the Management of the Open Data Initiative .....	1917
g)	Perceived Social Justice Implications of Open Data .....	1918
3.	<i>Implications of Stakeholder Assessment</i> .....	1919
B.	QUALITATIVE ASSESSMENT II: THE CITY .....	1920
1.	<i>The City of Seattle as a Case for Study</i> .....	1920
2.	<i>Selected Departments: A Sample Size of Eight</i> .....	1921
a)	The Department of Information Technology .....	1922
b)	The Department of Planning and Development .....	1924
c)	Finance and Administrative Services.....	1924

	d)	Seattle City Light .....	1927
	e)	Department of Transportation .....	1928
	f)	Police Department.....	1929
	g)	Parks and Recreation .....	1931
	h)	Fire Department .....	1931
	3.	<i>Analysis</i> .....	1932
C.		TECHNICAL ASSESSMENT: OPEN DATA ANALYSIS .....	1934
	1.	<i>The Problem of Cumulative Risk of Re-Identification</i> .....	1934
	2.	<i>A Proposed Method of Ex Ante Evaluation</i> .....	1936
	3.	<i>Potential Join Strategies</i> .....	1938
	4.	<i>Analysis and Results</i> .....	1939
	a)	Joins Using Exact and Flexible Matching Strategies.....	1940
	b)	The Special Relationship Between Municipalities and Spatial Data .....	1941
	c)	Attributes on a Continuum of Personalization .....	1944
	d)	One Simple Example of a Profile.....	1945
	5.	<i>Open Data Assessment in Sum</i> .....	1946
D.		LEGAL ASSESSMENT: VENDOR CONTRACTS.....	1947
	1.	<i>Privacy</i> .....	1948
	2.	<i>Security</i> .....	1951
	3.	<i>Analysis</i> .....	1953
IV.		RECOMMENDATIONS .....	1954
A.		INVENTORY DATA ASSETS .....	1954
B.		REQUIRE EACH UNIT TO DEVELOP AND SUBMIT DATA POLICIES.....	1956
C.		ESTABLISH NESTED GOVERNANCE STRUCTURE.....	1958
D.		ESTABLISH AND DISSEMINATE EX ANTE PROTOCOLS FOR PUSH, PULL, AND SPILL.....	1960
E.		CONDUCT PUBLIC RECORDS AUDIT AND TRAINING .....	1960
F.		EXPLORE CONDITIONED ACCESS OF MUNICIPAL DATA.....	1961
G.		DEVELOP STANDARD VENDOR AGREEMENT .....	1963
V.		FUTURE WORK.....	1965

## I. INTRODUCTION

Cities hold considerable information, including details about the daily lives of residents and employees, maps of critical infrastructure, and records of internal deliberations. Cities are beginning to realize that this information has economic and civic value. The responsible release of city information can result in greater efficiency and innovation in the public and private sector. New services are cropping up that leverage open city data to great effect.<sup>1</sup> Activist groups and residents are also placing increasing pressure on state and local government to be more transparent.

There has been little research into the growing area of municipal open data.<sup>2</sup> Cities are beginning to open their data in a way that has never been seen before, and these releases may raise privacy concerns. Scholarly and media attention has focused at the federal level toward the activities of the National Security Agency (NSA), the Federal Trade Commission (FTC), and the White House.<sup>3</sup> Despite the attention given to federal agencies, most personally-identifiable data is collected much closer to home, by the governments of the cities where we live, work, and play.<sup>4</sup>

---

1. See, e.g., Kathleen Hickey, *AppStore Gives Governments Access to Municipal Apps*, GCN (June 4, 2014), <http://gcn.com/articles/2014/06/04/granicus-appstore.aspx>; Angus Loten, *Entrepreneurs Shape Free Data into Money*, WALL ST. J., Jan. 9 2014; Jason Slotkin, *City Living: There's an App for That*, COMPUTERWORLD (Jan 11, 2013), <http://www.computerworld.com/article/2494114/mobile-wireless/city-living--there-s-an-app-for-that.html>; Geoffrey A. Fowler, *Apps Pave Way for City Services*, WALL ST. J. (Nov. 18, 2010), <http://www.wsj.com/articles/SB10001424052748704658204575611143577864882>.

2. For example, Maxat Kassen has observed:

[I]t is not yet clear how the potential of the open data concept can be realized at the local level as there has been no analysis of current projects so far. The concept is still in its infancy, and in fact it gained a political meaning primarily after the launch of the official U.S. government data portal in 2009. Later, similar data projects were initiated at the local level.

Maxat Kassen, *A promising phenomenon of open data: A case study of the Chicago open data project*, 30 GOV'T INFO. Q. 508, 509 (2013); see also Anneke Zuiderwijk & Marijn Janssen, *Open Data Policies, Their Implementation and Impact: A Framework for Comparison*, 31 GOV'T INFO. Q. 17, 17 (2014) (“[V]ery little systematic and structured research has been done on the issues that are covered by open data policies, their intent and actual impact. Furthermore, no suitable framework for comparing open data policies is available.”). As recently as 2011, the International City/County Management Agency national survey of e-Government did not include questions on open data. Donald F. Norris & Christopher G. Reddick, *Local E-Government in the United States: Transformation or Incremental Change?*, 73 PUB. ADMIN. REV. 165–175.

3. E.g., DANIEL J. SOLOVE, NOTHING TO HIDE: THE FALSE TRADEOFF BETWEEN PRIVACY AND SECURITY (2011); Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

4. See generally Bill Schrier, *Chapter 28: Toads on the Road to Open Government Data*, in OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN

This Article is a cross-disciplinary assessment of an open municipal government system. We are a team of researchers in law, computer science, information science, and urban planning that worked hand-in-hand with the City of Seattle, Washington to understand its current procedures around data processing from each of our disciplinary perspectives. Based on this empirical work, we have generated a set of recommendations to help the city manage risk latent in opening its data.

Seattle makes for a great case study. With a population of 650,000 and growing rapidly, Seattle is mid-sized, but not so enormous as to be unwieldy. It is a highly educated, technically savvy city and is often highly ranked among its peers on measures of innovation, creativity, and technology.<sup>5</sup> Seattle was one of the first cities to embrace an open data initiative.<sup>6</sup> Its leadership has publicly stated a need to achieve a balance between privacy and transparency.<sup>7</sup> During our research, we found encouraging signs in what Seattle is already doing and its willingness to adopt best practices, and identified areas for additional improvement.

#### A. THE MUNICIPALITY IN FOCUS

Municipalities govern a wide array of activities, from police services to building permits to parks and recreational services and facilities. City governments collect and process large amounts of information to support these activities, often with the help of third party contractors. Some of this data is confidential, requiring special handling for security purposes, while other is not confidential, but nevertheless contains sensitive details about residents and employees. If taken out of context or made publicly available, this data could bring about harms to privacy or social equity.

Rapid technological changes pose significant complications for municipalities seeking to govern data in the public interest. Municipalities are eager to become “smart cities” by adopting information technologies

---

PRACTICE 305, 305–313 (Daniel Lathrop & Laurel Ruma, eds., 2010); Kassen, *supra* note 2, at 509; Peter Conradie & Sunil Choenni, *On the barriers for local government releasing open data*, 31 GOV'T INFO. Q. S10, S10–17 (2014).

5. *E.g.*, Boyd Cohen, *The 10 Smartest Cities In North America*, CO.EXIST (Nov. 14, 2013, 7:08 AM), <http://www.fastcoexist.com/3021592/the-10-smartest-cities-in-north-america>.

6. Press Release, Socrata, Inc., Socrata Strengthens Open Data Market Leadership (Jun. 28, 2011), <http://www.socrata.com/newsroom-article/socrata-strengthens-open-data-market-leadership>.

7. Press Release, City of Seattle Office of the Mayor, City of Seattle Launches Digital Privacy Initiative (Nov. 3, 2014), <http://murray.seattle.gov/city-of-seattle-launches-digital-privacy-initiative>.

that promise more effective and efficient delivery of services.<sup>8</sup> Ubiquitous computing includes mobile micro-video cameras, utility meters that discern the use of appliances, and technologies for detecting and tracking residents' whereabouts, energy use, and other information. Each of these technologies has the potential to create real-time, continuous data feeds. As the technologies of data collection, processing, and storage become ever more advanced and potentially intrusive, local governments face the challenge of adapting policies and guidance about privacy and social equity to changing circumstances. In the absence of clear criteria and procedures, municipal agents may resort to ad hoc decision-making. In a federated system of governance, the cumulative implications of multiple data releases may have consequences not anticipated by any individual unit, including the ability to reconstruct the identity of an anonymous resident.

The data generated by municipalities is of interest to many commercial entities, which seek to use the data for purposes that are not necessarily aligned with the public interest. In March 2014, the FTC published a report introducing the data-broker industry, which is built around the collecting, processing and reselling of data about individuals.<sup>9</sup> Brokers aggregate data from public and private sources, index the data into detailed profiles of persons, households, and neighborhoods, and sell it to private and public buyers. Eight of the nine data brokers participating in the FTC study reportedly relied on information supplied by government to identify and profile individuals.<sup>10</sup>

## B. PURPOSE, THEMES, AND CONTENT

Our research explored both the mechanisms and consequences of municipal data releases. Our results provide a snapshot of activities and their

---

8. See generally Michael Batty, *Smart Cities, Big Data*, 39 ENV'T & PLAN. B: PLAN. & DESIGN, 191 (2012); Rob Kitchin, *The Real-Time City? Big Data And Smart Urbanism*, 79 GEOJOURNAL 1 (2014); Mike Weston, 'Smart Cities' Will Know Everything About You: How Can Marketers Cash In Without Becoming Enemies of the People?, WALL ST. J., July 12, 2015, <http://www.wsj.com/articles/smart-cities-will-know-everything-about-you-1436740596>. Weston writes:

[M]unicipalities and governments across the world are pledging billions to create "smart cities"—urban areas covered with Internet-connected devices that control citywide systems, such as transit, and collect data. Although the details can vary, the basic goal is to create super-efficient infrastructure, aid urban planning and improve the well-being of the populace.

*Id.*

9. FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY (2014), <https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>.

10. *Id.* at 15.

potential implications in a city that is striving to reap the benefits and avoid the pitfalls of data release.

*Cities share data in three basic ways: push, pull, and spill.* Cities “push” data when they publish databases through online or other portals. Residents and others “pull” data out of the city with public records requests. And cities “spill” data, through accidental exposure, malicious data breach, and the distribution of data by vendors, contractors, and partners. We use the push, pull, and spill taxonomy as a unifying theme throughout our analysis and recommendations.

Whether pushed, pulled, or spilled, the release of municipal data has many consequences. Three questions guided our exploration of the consequences of municipal data releases. Does the availability of open data increase public trust in the effective and efficient delivery of public services? Under what technological, legal, and other circumstances can municipalities govern the release of open data to meet the public need for privacy? What harms could municipal open data lead to, including issues of disparate racial or social impact, physical insecurity, or harm to consumers or the marketplace? We approach these questions across multiple methods and sections of this Article.

The rest of the Article proceeds as follows: We discuss our specific approach to investigating the city’s use of municipal data in Part II. Part III summarizes our findings. Part IV consists of seven recommendations for Seattle—and other cities interested in improving open data practices. We recommend: (1) conducting an inventory of data assets, (2) requesting each department to submit a data management plan, (3) establishing nested governance structures to deal with issues as they arise, (4) establishing ex ante and ex post protocols for push, pull, and spill, (5) conducting an audit and training around public records requests, (6) exploring the prospect of conditioned access to some city data, and (7) developing a standardized model contract for data vendors. We understand that Seattle is actively pursuing some or all of these recommendations even as of this writing. Finally, the Article closes with Part V outlining future work suggested by our analysis and findings.

## II. OUR APPROACH

There is little empirical work on municipal open data practices to date. However, exploratory research is not without guideposts. A sophisticated and expanding literature investigates the private sector’s use of information technology. This literature builds theoretical and empirical accounts and examines how those uses may compromise social norms and features of the economy; features that are prefaced upon the privacy of personal

information, racial and social equity, and the preservation of the public trust in digital or online transactions.<sup>11</sup> This Article seeks to begin a similar line of research aimed at the public sector, starting with municipalities. As subjects of research, municipalities are recent entrants into an ongoing, multidisciplinary conversation about the benefits and pitfalls of data collection, use, release, retention, commercialization, and security. This characterization is especially apt when the aim of research is to orient policy to the public interest.

As the subject of this particular study is municipal open data, we focus on the release of data by or from municipalities.<sup>12</sup> The push, pull, and spill taxonomy assisted us in designing research that would explore current practices while highlighting the potential future effects of such practices on public trust, privacy, and social equity. This required a mixture of research methods, each suited to a likely area of contest or hazard.

Our research methods and findings are described in four parts:

- **Qualitative Assessment 1—Key Stakeholders:**

We begin with a sense of the hopes and concerns of the parties affected by municipal practices. For this, we carried out focus groups on the topic of pushed, pulled and spilled municipal data, with several types of key stakeholders in the Seattle community. We relay our findings.

- **Qualitative Assessment 2—The City:**

We then discuss how Seattle itself handles data. We conducted interviews with city personnel involved in the release of data. Interviews spanned push, pull and spill: the intended purpose and use of open data by departments, the circumstances of public disclosure requests, and the involvement of departments in

---

11. See, e.g., Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security: Myths and Fallacies of “Personally Identifiable Information,”* 53 COMM. ACM 24 (2010), [https://www.cs.utexas.edu/~shmat/shmat\\_cacm10.pdf](https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf); Alessandro Acquisti & Jens Grossklags, *Privacy and Rationality in Individual Decision Making*, 3 IEEE SECURITY & PRIVACY, 26 (2005); Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010); Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. 995 (2014); Chris Jay Hoofnagle & Jan Whittington, *Free: Accounting for the Costs of the Internet’s Most Popular Price*, 61 UCLA L. REV. 606 (2014).

12. Other stages in the lifecycle of data matter and, though not central to this study, are just as worthy of research. The results of this study suggest promising future avenues for research in these related areas, including, for example, the potential for upstream decisions about collection and retention to be predicated on the downstream effectiveness of policies restricting the uses of data.

contracts with third parties for information-intensive services. The results indicate the types of data collected and used, the incentives that departments have to release datasets (or not), and the ways in which releases are modified to preserve privacy and social justice.

- **Technical Assessment—Open Data Analysis:**

We conducted technical analyses of the datasets already pushed to the City’s open data portal in order to understand how the City uses the portal and to investigate the extent to which the City’s current practices could potentially compromise privacy and social justice.

- **Legal Assessment—Vendor Contracts:**

Having identified, in departmental interviews, many contracts with third parties involving sensitive or confidential data about residents of the city, we examined these contracts for the kinds of safeguards one might expect in order to prevent, for example, unauthorized spills of this data.

As a collection of exploratory assessments, these research activities provide a broad array of insights into the role of the municipality in the release of data.

### III. FINDINGS

This part of the Article presents extensive findings on how a city generates and releases municipal data. This is a vast area for research. As other authors have explained, government departments are created to perform services that markets do not or should not provide, or are difficult or impossible for residents to provide for themselves.<sup>13</sup> For example, municipalities organize to provide regulatory functions to curb the many ways in which the for-profit, self-interested incentive structure of the private sector will “as if by an invisible hand” lead markets to fail to serve the public interest.<sup>14</sup> Within their jurisdiction, municipalities operate monopoly or monopolistic markets for several goods (e.g., water, electricity, roads, lighting), which are often provided through contracts with firms on behalf of residents. In negotiating these contracts, municipalities have

---

13. See Shrier, *supra* note 4, at 311.

14. See *id.* (quoting ADAM SMITH, AN INQUIRY INTO THE NATURE AND CAUSES OF THE WEALTH OF THE NATIONS 423 (Edwin Cannan ed., 1937) (1776)).

substantial leverage on the public's behalf, reducing the transaction costs that would have accrued if members of the public were left to organize and bargain on their own.<sup>15</sup> This bargaining power makes cities powerful market players—an untapped source of influence over privacy and security policy, as we discuss below. Municipalities also provide intergovernmental coordination: the geospatial area or jurisdiction of any given municipality is layered with the jurisdictions of several other governmental entities (e.g., special districts, counties, states, and the federal government). With such eclectic aims, municipalities can appear to be labyrinths of data production and release, bewildering in their complexity.

As a consequence of the enormity of the research task—as well as the inherent subjectivity in terms such as “open” or sensitive—we were forced to make certain assumptions and choices that we try to highlight through our findings. We also lay out an agenda for future work that reflects the realization that there is much more to do. Nevertheless, we attempted to convey and engage with both the breadth and depth of city data in our analysis.

Unlike physical assets, in Seattle as in many other cities, there is no central catalog of datasets and metadata. This research was conducted in partnership with the City of Seattle. The participation of departments in interviews and in the collection of key documents was critical to the success of this research in depicting, *in situ*, the governance of municipal open data.

#### A. QUALITATIVE ASSESSMENT I: KEY STAKEHOLDERS

Though our subject is municipal data, our backdrop is the people it affects. This section discusses our qualitative analysis of stakeholders' perceptions of open municipal data, particularly its downstream impacts. We understand that cities want to be responsive to their constituents, and

---

15. As Ronald Coase explains, illustrating with the case of the harmful effects suffered by many from the smoke exhaust of a factory:

[D]irect governmental regulation will not necessarily give better results than leaving the problem to be solved by the market or the firm. But equally there is no reason why, on occasion, such governmental administrative regulation should not lead to an improvement in economic efficiency. This would seem particularly likely when, as is normally the case with the smoke nuisance, a large number of people are involved and in which therefore the costs of handling the problem through the market or the firm may be high.

Ronald H. Coase, *The Problem of Social Cost*, 3 J.L. & ECON. 1, 18 (1960). On the application of Coase's theory to privacy harm through transactions with personal information, see generally, Hoofnagle & Whittington, *supra* note 11, and Jan Whittington & Chris Jay Hoofnagle, *Unpacking Privacy's Price*, 90 N.C. L. REV. 1327, 1331 n.9 (2012).

we endeavored to gain a sense of the hopes and fears of residents and others around open municipal data. We designed the research question for this component to be open-ended and as inclusive as possible of the range of issues that stakeholders may find relevant to the initiative. Through focus groups and interviews, we asked users for their hopes, concerns, and expectations for Seattle's open data initiative.

1. *Methods: Data Collection and Analysis*

a) Research Design and Sampling

The data collection for this study included the following stakeholder groups: (1) Seattle residents in general, (2) civic hackers, (3) privacy activists, (4) city employees, (5) an academic, (6) a legal advocate, and (7) industry representatives.<sup>16</sup> Our hope was to talk to those who directly use or would potentially use open municipal data, as well as those who work on closely related issues. Thus, with the exception of the group of "residents in general," respondents were largely familiar with the topic at the time of the focus groups and interviews.<sup>17</sup>

Seattle's local tech economy offers unique access to major industrial players, tech hobbyists, and activists. Data collection for this study was conducted with these existing organizations. For example, the "civic hackers" focus group was conducted with a local hobbyists group which meets weekly to build apps of local interest using open data. The focus group with privacy activists was conducted with members of a community activist organization focused on privacy issues, like the use of police surveillance cameras. The four industry representatives interviewed came from relevant departments in three large local corporations.

Most sampling for the study was purposive, based on respondent membership in relevant organizations or interest in the study.<sup>18</sup> Civic hackers, privacy activists, the legal advocate, academic, and industry representatives were contacted directly for their relevance to the study.

---

16. We adopt the Value Sensitive Design definition of stakeholders: "Direct stakeholders refer to parties—individuals or organizations—who interact directly with the computer system or its output. Indirect stakeholders refer to all other parties who are affected by the use of the system. Often, indirect stakeholders are ignored in the design process." Batya Friedman, Peter H. Kahn, Jr. & Alan Borning, *Value Sensitive Design and Information Systems*, in *EARLY ENGAGEMENT AND NEW TECHNOLOGIES: OPENING UP THE LABORATORY* 55, 73 (2013).

17. We used a focus group format to collect data from the first four stakeholder types listed. Due to scheduling constraints, data from a legal advocate, academic, and industry representatives was based on interviews.

18. As part of the University of Washington Institutional Review Board (IRB) approval for this study, demographic information about respondents was not collected.

Members of the general public were recruited via fliers and Craigslist.<sup>19</sup> Our hope for the city employees focus group was to speak with workers on the “front-line”—police, fire, waste management, and others who drive fleet vehicles; constraints within the city made this infeasible. The city employees who participated were largely administrative staff; nevertheless, this group was more sensitive to potential privacy issues than we had expected.

b) Data Collection

Data collection for this study was based on focus groups and interviews. The focus group format was piloted twice to make it more neutral. Each focus group had 7–10 members and lasted 60–120 minutes. We used this format for residents, privacy advocates, civic hackers, and city employees.<sup>20</sup> Focus groups are well-suited for understanding unobservable phenomena like attitudes.<sup>21</sup> As a method, focus groups present a risk of respondent bias and group-think; our research design took measures to minimize these risks.<sup>22</sup>

Focus groups began with a 10-minute introduction from the moderator covering relevant background information. The moderator introduced the city’s open data portal, the types of data currently available on it, and data types that the city has made available. The moderator introduced the Washington State Public Records Act (PRA), and its strong value on government transparency.<sup>23</sup> The PRA is a state law that establishes broad rights for state residents to request public records. It is intended to promote government transparency and accountability. The moderator explained that while the PRA requires the reactive release of data in light of a public disclosure request, open data is proactively released and not mandated. The presentation discussed how data is anonymized by removing its identifying

---

19. This group was compensated \$15 for their time. No other respondent was compensated. Perhaps because of this means of recruitment, respondents for the general public group happened to be people experiencing instability in employment and housing.

20. In addition, we also interviewed four industry representatives, a legal advocate, and an academic.

21. For a detailed discussion of the strengths and weaknesses of focus groups as a research method see DAVID L. MORGAN, *FOCUS GROUPS AS QUALITATIVE RESEARCH* 13–17 (2d ed. 1996).

22. See Jenny Kitzinger, *Qualitative Research. Introducing Focus Groups*, 311 *BRIT. MED. J.*, 299–30 (1995) (“The method is particularly useful for exploring people’s knowledge and experiences and can be used to examine not only what people think but how they think and why they think that way.”). See generally Jenny Kitzinger, *The Methodology of Focus Groups: The Importance of Interaction Between Research Participants*, 16 *SOC. HEALTH & ILLNESS* 103 (1994).

23. See WASH. REV. CODE § 42.56 (2011) (Public Records Act).

attributes, and under what circumstances data subjects may be re-identified, if any. Focus groups were conducted with a minimal moderation approach.<sup>24</sup>

c) Data Analysis

Transcripts of the focus groups and interviews were analyzed via qualitative coding. The first round of coding used a priori codes based on our research questions. The second round of coding used open and axial coding, in keeping with a grounded theory approach.<sup>25</sup> Analysis was conducted using NVivo 10 qualitative data analysis software.<sup>26</sup> Using this tool, the researcher tags blocks of text with a theme. Based on these tags, the software creates a database of quotes indexed by theme and respondent group. Iterative, inductive coding was formalized as a coding manual, by which data analysis was standardized across respondent groups. In keeping with a grounded theoretic approach, the following results are closely derived from the data.

The results of the stakeholder analysis offered a range of perceptions on the downstream impact of open data. Due to the exploratory, open-ended nature of this study, the analysis covered a broad scope of hopes, concerns and expectations about who will use the data, and to what end. Issues related to public trust, privacy, race, and social justice were of core interest to this work. Additional topics, like safety, commercial actors, and legal issues also emerged in the analysis. In this section, we discuss results by theme, and offer a sense of the inter-group variation on a given issue.

---

24. Respondents were told that the central goal of the session was to hear as many of their hopes and concerns as possible. Three themes—public trust, privacy, and race and social justice—were of particular interest to this project. Rather than prompting these themes directly, the moderator waited to see if they arose naturally from the conversation. If any of these topics were not addressed, the moderator made a note of this, then directly addressed remaining themes at the end of the session.

25. Qualitative coding is an interpretive process of systematically analyzing a text to surface themes within it. A priori codes are themes that the researcher brings to the text. A grounded theory approach necessitates that these themes arise from the text itself. Open coding is the initial process of capturing each theme from a text; axial coding combines these open codes into groups. For background on these coding methods, see generally Juliet Corbin & Anselm Strauss, *Strategies for Qualitative Data Analysis*, in *BASICS OF QUALITATIVE RESEARCH: TECHNIQUES AND PROCEDURES FOR DEVELOPING GROUNDED THEORY* 85 (4th ed. 2014).

26. See *What is NVivo*, QSR INT'L, <http://www.qsrinternational.com/what-is-nvivo> (last visited Sept. 23, 2015).

## 2. Findings

### a) Effects of Open Data Initiative on Public Trust

Respondents' primary hope for open data was that it would increase transparency in government. Every group touched on this sentiment, although the form it took varied. This included hopes for greater transparency, the democratization of governance, and the hope to build a better society through data-driven policy decisions. Government accountability was of keen interest to those in five of the seven stakeholder groups. This was expressed in many forms, from oversight on police or prison guard actions, to residents fact-checking politicians by looking at the same raw data. Some groups, like the civic hackers, presented this hope with conviction: "Having the data be open is an incredible source of accountability. It is a key to democracy."<sup>27</sup> This group spoke in-depth about opportunities for widespread data-literacy, which was viewed as a key intermediate step to true accountability. Others, especially privacy advocates, and residents in general, held similar hopes while also more ambivalent; we outline these concerns further on.

### b) Economic Value Latent in Data

A commonly stated goal for open data is that it can bolster the local economy. Stakeholders—including industry representatives, privacy activists, and civic hackers—shared this goal. Some focused on ways open data can foster new companies and lead to more jobs, or allow existing companies to offer new products. Industry representatives were interested in ways that commercial actors improve the quality of data as they use it, and cited the potential for a "two-way pipe," by which companies could add value to the data—e.g., with real-time data feeds—and give it back to the city.<sup>28</sup> One industry representative said data could be used to target their marketing: "How do you find out which customers are heavy commuters? You just ask the city for all the tapes about license plates."<sup>29</sup> Privacy activists and hackers said that businesses could help interpret and make the data more usable to everyday people. However, one privacy activist thought that while analysis and usability was a valuable role for businesses, it constituted a public good that should not be delegated to private actors. Civic hackers were hopeful that open data could help smaller, more agile companies replace large firms in government procurement.

---

27. Focus Group, Civic Hacker Organization, in Seattle, Wash. (Feb. 12, 2015).

28. Telephone Interview, Industry Representative #2 (Mar. 27, 2015).

29. Telephone interview, industry representative #1 (Mar. 27, 2015).

c) City Management of Open Data Initiative

Stakeholders asserted a range of expectations for the city in how they proceed with the open data initiative. Every group stated that the data should be anonymized prior to release. In keeping with the spirit of the PRA, there was also a strong conviction that data held by government belonged to the public. The groups who most used this data, like industry representatives, privacy activists and civic hackers, had specific input for the way the data is and should be stored, accessed, formatted, licensed and released. These groups stated that the license terms under which the data was released should be clearer. The legal advocate and academic shared the expectation that the city should limit data collection, and limit its use beyond that for which it was collected. Despite potential risks, civic hackers and privacy advocates were profoundly opposed to the idea of access restrictions, fearing that they would be used against someone with legitimate interest in the data. Often, the scope of this conversation moved into one about the city as a data custodian: its data storage, retention, and deletion processes.

Multiple groups shared a sense of unease about the city's ability to prevent data spill.<sup>30</sup> This concern was echoed by members of the general public, who were acutely concerned about hacking and identity theft. Both industry and city employees said that the city's servers are regularly targeted by Chinese hackers and other international actors. As we discuss further on, both the general public and city employees were concerned that hacked data would be used to threaten critical infrastructure.

There was large variation within and between groups on the feasibility of use restrictions on the data, with an overall sense that restrictions would not be enforceable. Civic hackers and privacy activists noted the practical problems with governing uses of data once it is made open. The legal advocate pointed out that some forms of use restrictions would represent unconstitutional restraints of free speech. Even in the absence of formal use restrictions, industry representatives were sensitive to the way the public

---

30. One industry representative said:

They need to follow reasonable baseline data security practices, particularly if the city is going to be a repository of big data. And, if for-profit companies in the health-care sector, for example, have under-invested in data security, then it's a fair bit to say the IT systems of many municipal governments aren't where they should be either.

*Id.*

would react to different uses.<sup>31</sup> Public-facing organizations, as opposed to organizations that work business-to-business, were thought to use public feedback as a check on data uses. The legal advocate shared this sense, adding that data brokers and less visible actors are less responsive to norms around data use: “Is anyone really comfortable with the variety of awful things that have happened with commercial actors in this space—like companies creating extortion schemes by posting photos of people online that they get via public records?”<sup>32</sup> While use restrictions were generally deemed infeasible, this quote illustrates the ambivalence stakeholders expressed about unintended consequences of data release.

d) Privacy Interests in Open Data

Privacy implications of the open data initiative were a prominent feature in every conversation, with the exception of the civic hackers group. Some respondents among the general public and civic hackers asserted that “privacy is an illusion.”<sup>33</sup> Members of these groups strongly believed a data spill was liable to happen eventually. However, they were less concerned about privacy implications than they were that public outcry would slow the momentum of the open data initiative. Civic hackers framed concerns about privacy as important, but coeval with concerns about data inaccuracy and misinterpretation. Overall, this group shared an impetus to get “more eyes on more data”<sup>34</sup>—data in anonymized form. Some respondents in the privacy activist group shared the civic hackers’ confidence that data

---

31. One representative said, “We’re very conscious of ethics and big data, civil rights and big data, and trying to be really thoughtful about how we combine data so that it isn’t used in bad ways or identifies people.” Telephone Interview, Industry Representative #3 (Mar. 27, 2015). Similarly, another industry representative said:

[I]t could be useful for commercial benefit if you’re doing that in a de-identified or aggregated way, and that shouldn’t be a problem. If you’re doing it in a personally identifiable way—so the people can add factors to your behavioral profile—that’s probably going to rub people the wrong way.

Telephone Interview, Industry Representative #1 (Mar. 27, 2015).

32. Interview, Legal Advocate, in Seattle, Wash. (Feb. 19, 2015).

33. Focus Group, Civic Hacker Organization, in Seattle, Wash. (Feb. 12, 2015). One civic hacker said:

I think that banks and private health care are a much bigger concern for privacy problems than the government; they’re a lot more focused. [Governments have] bits and pieces of data all over the place, you’d have to really want to aggregate that stuff in order to really drill down in somebody’s privacy.

*Id.*

34. *Id.*

anonymization processes are resilient to reverse re-identification. Members of the general public and the legal advocate were less confident that data anonymization could protect individuals.

Other stakeholders had more acute privacy concerns. There was a general sense that the city had sensitive data. A privacy advocate said, “I fear the efforts to make data available about the government actually makes data available about the public.”<sup>35</sup> The category of what information is or should be “private” varied between groups. Members of the general public framed private data as social security numbers and information related to financial status (e.g., credit rating). An industry representative and civic hackers emphasized that locational data would be a privacy concern, if released in a granular way. The legal advocate favored an approach that would scrutinize any data type as one piece of a larger mosaic: “If it’s a sufficient analysis, it’s also going to take into account whether this information, when correlated with other data that is available, presents harms.”<sup>36</sup> The legal advocate spoke to ways that data could be re-identified; thus, he said that entire record types should be considered sensitive (e.g., police video) and exempt from proactive release or most forms of public records request.

City employees’ discussion of what constitutes private information was broader than that of other groups, due in part to the large amount of information the city has in their personnel files. Employees described the different standards of privacy that applied to them as public employees. They recalled the shock of adjusting to having their salaries posted publicly. Members of the group were unaware of whether certain data types were protected from public records request under the PRA, for example, home address, employee benefits, and retirement information. These respondents were also very concerned about the release of insurance information such as the identity of their dependents or other family members.

Multiple respondents within all groups mentioned specific segments of the population they perceived as having special privacy interests. Several groups, including the general public and civic hackers, mentioned the special interests of children and the elderly. One privacy activist said:

It’s a really privileged position to be able to say that everything should be open. People with experiences of different kinds of abuse have had to build hiding into their cultural identity—open is not just going to work for them.<sup>37</sup>

---

35. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

36. Interview, Legal Advocate, in Seattle, Wash. (Feb. 19, 2015).

37. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

Safety concerns were the primary reason cited for these special privacy interests.

e) Safety Risks Latent in Data

Concerns about safety were more widely held than we had expected, and came up in conversations with every group. Respondents were concerned about the safety of vulnerable populations. There was concern that children, elderly people, and victims of previous crimes would be specifically targeted by criminals seeking to assault or con them. These concerns were brought up widely, in five out of seven groups. The nature of government services means that those in need will be especially present in the data. One City employee pointed out ways that police officers' route information reveals domestic violence: "you can find safe houses, individuals that are maybe victims that are being involved in their processes and response patterns."<sup>38</sup> Privacy activists noted that governments also have data on foster children and those in child protective services.

Multiple respondent groups were concerned with the safety implications for City employees. First responders were perceived to be at risk of vigilante justice. A privacy activist said:

People have tried to find out where cops live so they can go to their houses and do stuff to them. Cops still have personal rights and personal privacy rights and stuff too, even though we would default to thinking that they don't go out of their way to respect our own.<sup>39</sup>

This concern for officers' safety co-existed with the respondent's other attitudes about police. City employees even referred to a past PRA request for police officers' home addresses that had been granted. They noted that this incident had led the fire department to take greater precautions with the kinds of identifying information it included in its reports.<sup>40</sup> City employees also raised the possibility that public data could be used to derive route patterns, which could be used by criminals to target officers on their daily routine.

---

38. Focus Group, City Employees, in Seattle, Wash. (Mar. 9, 2015).

39. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

40. One exchange in this group illustrates these concerns:

You might get incident information, but you're not going to have the firefighters' names because then they're easily looked up. They're at Station X, OK—you can see shift details and stuff, so we have to be smart about it. Especially the kinds of shifts firefighters are on—they have to leave their families . . . They're on 24 hours.

Focus Group, City Employees, in Seattle, Wash. (Mar. 9, 2015).

City employees were also concerned about their safety. Some responded that they felt they could be targeted because of their race or sexual orientation; one person described a city department's LGBT group meeting wherein another city employee tried to use the PRA to request the names of all attendees. The same person reported feeling outed when trying to change his or her official marital status.<sup>41</sup> Other respondents felt personally exposed by ways that public records are indexed and searchable on Google.

Safety risks were perceived to implicate not only individuals, but larger domestic security concerns. Members of the general public, industry representatives, and city employees referred to the potential for open data to be used to target critical infrastructure. This risk was framed as applying to physical infrastructures, like the power grid, as well as servers and other digital assets. To the extent that open data could be used to derive first response patterns, city employees were concerned that this information would be used to divert public safety officers from a planned attack. The academic cited a counterexample of the public safety utility of open data, especially public health concerns like vaccine and disease status.

f) Lack of Public Trust in the Management of the Open Data Initiative

Despite these risks, multiple stakeholder groups were concerned that the government would not open enough data. Civic hackers, industry representatives, privacy activists, and members of the general public shared a concern that open data efforts would fall short of its promise if very little data were released. Members of the general public and privacy activist groups shared a sense that those in city government would selectively record or release data to protect their own image. One privacy activist said, "If the city . . . maintains the ability to selectively refrain from publishing portions of that data, then we're not a whole lot better off than if they just weren't publishing in the first place."<sup>42</sup> Respondents in the civic hackers group and

---

41. This individual responded:

It doesn't feel safe to me at all. My being, you know as a, being married, I had to contact a lot of people to get my status change in the city. They didn't, you know, so then I'm thinking okay, let's advertise it even more to everybody. I was certainly in my right so I'm going to do it, but it's pretty public. If I wanted to not tell people I was gay, it would have been impossible because everybody has access to it.

*Id.*

42. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015). A member of the Seattle residents group said, "This is just something they're doing to appease the general public because there's an outcry in America. But the police is going to be the police . . . as soon as they get some information they don't want to be publicized,

privacy activists were concerned that government actors could edit data, and raised the importance of using metadata or a data signature or hash that would verify its authenticity. While the responses of the general public and privacy activists exhibited low trust in government, civic hackers were more interested with issues of data quality.

Other groups worried that the promise of open data might become encumbered. One industry representative thought agencies might lose sight of the larger goals surrounding open data. He feared open data would become “a compliance exercise where the agencies and the cities will all do whatever they have to [do in order to] stop being bothered about it anymore.”<sup>43</sup> This respondent spoke from a sense that unambitious management of the data would pose a missed opportunity. Both civic hackers and city employees noted that governments feared exposing themselves to liability from data release; for the civic hackers, liability and related concerns were framed as barriers to progress.

#### g) Perceived Social Justice Implications of Open Data

Respondents perceived open data as having promise for social justice issues. Half of the groups explicitly mentioned “social justice” issues without prompting. Even when not referred to explicitly, the implication of open data on social justice issues was present in respondents’ ideas about government accountability for misconduct. Other references to social justice included the possibility of communities using data to advocate for themselves (civic hackers), data-driven policy (general public and civic hackers), and crowdsourced service requests (e.g., potholes, streetlight reports) (industry representatives). While some in the general public group felt that open data would have positive and incremental social justice implications, one person thought that little would happen in this vein: “I think the reality of it is, it’s not going to really affect anybody that’s down and out anyway in Washington State, it’s only going to affect the . . . powers-that-be anyway.”<sup>44</sup> Racial minorities within the general public group expressed a sense that open data would not be put to work on their behalf.

Other groups raised concerns that open data could have negative racial and social justice implications. Many of these were related to the potential that commercial uses of the data would have a disproportionate impact on marginalized communities. One member of the privacy activists said, “I fear

---

there’s going to be a glitch in it.” Focus Group, General Public, in Seattle, Wash. (Mar. 19, 2015).

43. Telephone Interview, Industry Representative #4 (Apr. 6, 2015).

44. Focus Group, General Public, in Seattle, Wash. (Mar. 19, 2015).

that it would be used to lower property values, redline insurance, et cetera, in neighborhoods with high crime rates rather than addressing those issues. I'm worried that data about precincts where people don't vote much could lead politicians to write them off."<sup>45</sup> A member of the general public group spoke to the ways that data, once open, is copied and persists:

The information they put on there is a detriment to me because I've been trying to get, well I just got out. I was released from a penitentiary and I've been trying to get work and anytime they do a background check it's bringing up shit from like 1996. This is 2015.<sup>46</sup>

Taken together, these responses highlight how uses for open data could reify existing social marginalization.

### 3. *Implications of Stakeholder Assessment*

The open-ended nature of the qualitative stakeholder assessment resulted in some findings that we might have expected, some opinions that were more widely shared than we would have expected, and some surprises. For the purpose of our recommendations, we foreground the following results: (1) Multiple groups expressed concern regarding privacy risks latent in the data, especially to vulnerable and marginalized populations and city employees. Not all stakeholders were confident that anonymization would be enough to protect those listed in the data, although each stakeholder listed strong anonymization as an expectation for the city. (2) Stakeholder groups spoke to positive economic impacts from commercial uses of the data, but drew a clear line between these uses and those that were considered overly intrusive. Members of the general public were aware of threats to privacy from data brokers, which the research team did not expect. (3) City employees did not know what aspects of their personal data were protected, and they did not feel safe. (4) In thinking about open data, many groups spoke more broadly about issues of data custodianship; in their eyes the city's responsibility to protect its data and to open it intertwined. (5) Stakeholders were not clear about the terms under which data was released, and asked for data licensing, with more clear terms. (6) Respondents were concerned about ways that governments might prevent data release to protect itself, or might treat different data requestors differently. Our recommendations were shaped in part by the application of these findings.

---

45. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

46. Focus Group, General Public, in Seattle, Wash. (Mar. 19, 2015).

## B. QUALITATIVE ASSESSMENT II: THE CITY

Having generated some context for our discussion by connecting with residents and other stakeholders, we turn to a discussion of how the City of Seattle actually processes and shares data. This section discusses the findings of interviews with city departments relevant to municipal data management and release.

### 1. *The City of Seattle as a Case for Study*

One underlying premise of this research is the tension or conflict between the adoption of “smart city” technology and the protection of privacy and fairness for the individuals and groups who generate the data. In this respect, recent events have made Seattle an ideal case for study. On February 3, 2015, the City of Seattle formulated and adopted a set of privacy principles, which will guide the actions the city takes when collecting and using personal information. Central to the principles is the following policy statement: “We work to find a fair balance between gathering information to provide needed services and protecting the public’s privacy.”<sup>47</sup> The six privacy principles adopted speak to the importance of keeping personal information private when collecting it, storing and using only what is needed for city services, and being accountable for “managing your personal information in a manner that is consistent with our commitments and as required by law.”<sup>48</sup> Where possible, the City also commits to updating information to be accurate, and notifying citizens on how information is used.

Many Seattle departments have adopted or contracted for the use of various smart city technologies to improve the efficiency and effectiveness of public services. Smart cities have been defined according to their use of large-scale sensor networks to improve the provision of city services.<sup>49</sup> As Rob Kitchin explains,

The notion of a ‘smart city’ refers to the increasing extent to which urban places are composed of ‘everyware’; that is, pervasive and ubiquitous computing and digitally instrumented devices built into the very fabric of urban environments (e.g., fixed and wireless telecom networks, digitally controlled utility services and transport infrastructure, sensor and camera networks, building

---

47. CITY OF SEATTLE, PRIVACY PRINCIPLES, <http://www.seattle.gov/Documents/Departments/InformationTechnology/City-of-Seattle-Privacy-Principles-FINAL.pdf>. Disclosure: One of us assisted Seattle in its formulation of privacy principles through his participation in an advisory board.

48. *Id.*

49. Kitchin, *supra* note 8, at 1–2.

management systems, and so on) that are used to monitor, manage and regulate city flows and processes, often in real-time, and mobile computing (e.g., smart phones) used by many urban citizens to engage with and navigate the city which themselves produce data about their users (such as location and activity).<sup>50</sup>

The adoption of these technologies amongst Seattle's departments, and the simultaneous adoption and development of citywide privacy principles, signify the tension that exists between the perceived role of the city as a custodian, consumer, and distributor of data about residents. Depending on the perspective one has, or rationale one adopts, the same categories of data may be considered either to be of value to the public—therefore warranting public distribution, or of value to the public—meaning it should be kept in a secure state with strict controls on access.

## 2. *Selected Departments: A Sample Size of Eight*

Like virtually all mid- to large-sized municipalities, the City of Seattle functions more as a federated system of departments than a hierarchy.<sup>51</sup> The open data portal in Seattle is the product of activities conducted by the Department of Information Technology, which oversees the third-party contractor who maintains the portal. However, each department in the City governs the data it generates with considerable autonomy.

With regards to the release of data, departments are also subject to many different rules and regulations, from both internal and external sources. The Washington PRA, however, applies to all departments.<sup>52</sup> Thus, many of the City's units are involved in the release of data.

The City of Seattle contains thirty-six departments and agencies.<sup>53</sup> Within this population, we selected eight to research: the Department of Information Technology; the Department of Planning and Development; Finance and Administrative Services; Seattle City Light; the Department of Transportation; the Police Department; Parks and Recreation; and the Fire Department. A few criteria, generally organized around the principles of maximizing internal variation and generalizability, guided our selection. In consultation with City staff, departments were selected to represent the variety of challenges and approaches cities face as data is pushed, pulled, and spilled. Most, but not all of the selected departments, are active users of the

---

50. *Id.* (internal citations omitted).

51. In comparison to private firms, municipalities appear to be very flat organizations. This is due in part to the sheer number of roles and responsibilities mandated for and by local government.

52. *See* WASH. REV. CODE § 42.56.010(1) (2014).

53. *See* *Departments and Agencies*, SEATTLE.GOV, <http://www.seattle.gov/city-departments/departments-and-agencies> (last visited June 23, 2015).

open data portal. Many, but not all, are undergoing rapid changes in data management due to the adoption of new information technology. Almost all govern at least some data that is understood to be either sensitive or confidential, though the characteristics of the data subjects and the attributes of those datasets differ considerably. This list includes the departments that receive the greatest demand for public disclosure requests, but also some that experience very few. They rely on a wide variety of third party contractors for information-intensive services.

Importantly, however, departments were selected to represent the variety of technologies and enriched information flows that are the hallmark of smart cities. For this purpose, we based selection on a rationale categorizing sensors and data subjects as “stationary” or “mobile.” Both a sensor and data subject can be stationary, as is the case with advanced meters with sensors that automatically record electrical or water use in the home or office. The sensitivity of this data is generally a function of its granularity over time. A sensor can be stationary while the subject of the data is mobile. This is the case in the study and provision of transportation services, which track the movements of data subjects. Both the sensor and data subject can be mobile. Video cameras hoisted on police patrol cars or pinned on the lapels of police officers’ uniforms are examples. This schema is useful for beginning to think about ways that information technology advances can result in the production of more sensitive data.

With the eight departments selected, in-person and telephone interviews were conducted with departmental personnel in various roles associated with the push, pull, and spill of municipal data.

a) The Department of Information Technology

Shortly after President Obama signed the 2009 Memorandum on Transparency and Open Government,<sup>54</sup> the start-up firm Socrata approached the Department of Information Technology about purchasing its services to support open data. After about a year of conversation, Seattle contracted with Socrata and began the process of selecting and examining datasets for release to an open data portal.<sup>55</sup>

In considering the publication of data, the Department of Information Technology uses a classification system with four levels:

**Public Information:** Public information can be or currently is released to the public. It does not need protection from

---

54. Transparency and Open Government, 74 Fed. Reg. 4685 (Jan. 26, 2009), [https://www.whitehouse.gov/the\\_press\\_office/TransparencyandOpenGovernment](https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment).

55. Interview, Department of Information Technology personnel, Seattle, Wash. (Jan. 21, 2015).

unauthorized disclosure, but does need integrity and availability protection controls. This would include general public information, published reference documents (within copyright restrictions), open source materials, approved promotional information and press releases.

**Sensitive Information:** Sensitive information may not be specifically protected from disclosure by law and is for official use only. Sensitive information is generally not released to the public unless specifically requested. Although most all of this information is subject to disclosure laws because of the City's status as a public entity, it still requires careful management and protection to ensure the integrity and obligations of the City's business operations and compliance requirements. It also includes data associated with internal email systems and City User account activity information.

**Confidential Information:** Confidential information is information that is specifically protected in all or in part from disclosure under the State of Washington Public Disclosure Laws. This could include certain personally identifiable information or vendor trade secrets.

**Confidential Information Requiring Special Handling:** Confidential information is specifically protected from disclosure by law and subject to strict handling requirements dictated by statutes, regulations, or legal agreements. Serious consequences could arise from unauthorized disclosure, such as threats to critical infrastructure, increased systems vulnerability and health and safety, or legal sanctions. Departments handling this category of information must demonstrate compliance with applicable statutes, regulatory requirements and legal agreements. Information in this category could include patient health records or student school records.<sup>56</sup>

Note that the first level pertains to data the City considers applicable for posting as open data (push). The second pertains to data that is subject to disclosure by request (pull). The last two levels pertain to confidential data for which City staff have "a legal reason to refuse public disclosure."<sup>57</sup>

On the incentives for releasing data, department personnel suggest that they try to save costs on public disclosure requests. The message that pushing data to an online portal may result in more efficient public

---

56. E-mail Communication, Department of Information Technology Personnel (Jun. 29, 2015).

57. Interview, Department of Information Technology Personnel, Seattle, Wash. (Jan. 21, 2015).

disclosure is reinforced by the PRA, which notes, “The internet provides for instant access to public records at a significantly reduced cost to the agency and the public. Agencies are encouraged to make commonly requested records available on agency web sites.”<sup>58</sup> Another rationale for municipal open data is the prospect of promoting economic or business growth in the city after the Great Recession. Importantly, department personnel also express hope that public open data has been anonymized properly. As they say, “how do you make a race car go faster? You give it better brakes.”<sup>59</sup>

b) The Department of Planning and Development

One of the early and active participants in the open data portal was the Department of Planning and Development.<sup>60</sup> Most city datasets that concern infrastructure do not pertain to critical infrastructure. Among the datasets made public by the Department are Geographic Information System (GIS) files that show plans, land use, zoning, critical areas, topography, vicinity to park property, landmarks, planning and permits. All permits for work done on private property are posted to the open data portal. Department personnel describe the postings as “complete,” and they can potentially include location, the property owner’s identity, and the work performed.

The Department of Planning and Development, like all departments contributing open data, is thought to be the “owner” of the data, and it is up to their discretion whether to participate. The rationale behind Planning and Development’s decision to participate is common to many departments that publicize data. Departments consider “the business case”: is this data subject to repeated public disclosure requests? Would the preemptive preparation and release of the data through the open data portal save time and resources when compared to responding to public disclosure requests?<sup>61</sup>

c) Finance and Administrative Services

In the first analysis of sensitive data for release to the open data portal, the Department of Information Technology worked with Finance and Administrative Services to assess the risk of making business license data publicly accessible. As explained in their risk analysis:

---

58. WASH. REV. CODE § 42.56.520, finding 2010 c 69 (2010).

59. Interview, Department of Information Technology Personnel, Seattle, Wash. (Jan. 21, 2015).

60. Interview, Department of Planning and Development Personnel, Seattle, Wash. (Jan. 14, 2015).

61. *Id.*

The Department of Finance and Administrative Services has developed a process for evaluating datasets against eight principles of open data and a risk analysis profile associated with publishing the data. The risk analysis defines who the final decision maker should be, and who will decide whether or not to publish the dataset.<sup>62</sup>

The principles the Departments referred to are the “8 Principles of Open Government Data,” formulated during a 2007 meeting convened by Tim O’Reilly, of O’Reilly Media, and Carl Malamud, of Public.Resource.Org, with sponsorship from the Sunlight Foundation, Google, and Yahoo.<sup>63</sup> The principles formulated by this group assert that open government data should be:

1. **Complete:** All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.
2. **Primary:** Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.
3. **Timely:** Data is made available as quickly as necessary to preserve the value of the data.
4. **Accessible:** Data is available to the widest range of users for the widest range of purposes.
5. **Machine Processable:** Data is reasonably structured to allow automated processing.
6. **Non-discriminatory:** Data is available to anyone, with no requirement of registration.
7. **Non-proprietary:** Data is available in a format over which no entity has exclusive control.
8. **License-free:** Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.<sup>64</sup>

---

62. City of Seattle Department of Finance and Administrative Services, Open Data Candidate Requirements and Risk Evaluation—Business License Data 3 (May 6, 2010), <http://dropbox.ashlock.us/opengov/seattle/Open%20Data%20Candidate%20Requirements%20and%20Risk%20Evaluation%20V1%209.docx> [hereinafter City of Seattle, Open Data Candidate Requirements].

63. *Open Government Data Principles*, PUBLIC.RESOURCE.ORG (Dec. 8, 2007), [https://public.resource.org/8\\_principles.html](https://public.resource.org/8_principles.html).

64. *Id.*

The Departments also added that customer service personnel responsible for constituent requests should be notified.<sup>65</sup> Seattle's risk analysis compared each data type in the business license dataset to each of these eight principles. Analysis proceeded field by field, noting which were to be excluded from release because they contained data for internal use only, of a personal nature, or data generated by the system (i.e., data that is only of use to those who operate the business registration system). For example, analysis of the data under the first of the eight principles revealed several fields that contained sensitive data, which should be excluded from release.<sup>66</sup>

The final recommendations focused on the potential legal risk if a data type were released. The Departments recommended publishing part of the dataset,<sup>67</sup> that is, publishing the dataset without mailing addresses and [personal] regulatory information. The analysis recommended that a subset of the data be extracted each month, and prepared for output to the open data portal.

The decision to publish the data was influenced by the perceived risks inherent in publication. Low-risk data could be published as is, while high-risk data required "too much data clean up" prior to publication.<sup>68</sup> Medium-risk datasets required exclusion of only certain fields. The business license dataset risk analysis concluded with the statement: "The risk for this dataset is rated at Medium, therefore the final approver for publishing this dataset to data.seattle.gov will be the [Finance and Administrative Services] director."<sup>69</sup>

While this example illustrates the reasoning and approach Seattle has taken toward releasing datasets on Socrata's platform, Financial and Administrative Services Department personnel note that the effort required to release secure data has escalated significantly.<sup>70</sup> The department is currently working in coordination with several other cities in the Puget Sound region on an initiative to convert all business and occupation (B&O) tax data to an online portal for processing payments and providing results to queries for tax information. While not open data in the same sense as the data pushed to the Socrata platform, this initiative also proposes to reduce

---

65. City of Seattle, Open Data Candidate Requirements, *supra* note 62, at 17.

66. *Id.* at 9.

67. *Id.* at 19.

68. *See, e.g., id.* at 9.

69. *Id.* at 18.

70. Interview, Department of Planning and Development personnel, Seattle, Wash. (Jan. 14, 2015).

costs to taxpayers by allowing secure, online payment and retrieval of tax information.

d) Seattle City Light

Seattle City Light is Seattle's publically owned electric power utility company. Currently, most of Seattle's residences are still outfitted with mechanical or relatively simple digital meters for reading and recording the rate of electricity consumption.<sup>71</sup> Seattle City Light employees take readings at the customer's residence or business location. This method delivers no more than six points of data per year, in sync with the utility's bi-monthly billing cycle.<sup>72</sup> However, technology in this sector has advanced rapidly, and Seattle's meter system is changing.

Seattle City Light has implemented three programs on a path toward smart metering. In 2008, the utility tried a pilot program with 457 meters that relied on cellular technology to provide daily, one-way, communication (from the customer's site to the utility).<sup>73</sup> Another estimated 6,000 meters, in places the utility describes as "hard to reach," are using radio frequency technology to signal usage to the utility.<sup>74</sup> For several years, the utility has also operated a program for customers who manage mid- to large-sized properties, providing continuous two-way communication through meters hooked up to phone lines. Referred to as Seattle Meter Watch, the program is part of a larger industry-led initiative, known as the Green Button Initiative. Since 2012, the Green Button initiative has been a White House-led effort to allow consumers to access detailed data about their electricity usage, and take advantage of online tools for saving money by managing their use. Seattle was the first utility in the nation to be certified under this initiative.<sup>75</sup>

As part of the utility's six-year Strategic Plan, Seattle City Light has begun to scale up the installation of advanced meters. Unlike the city's mechanical meters, which are simply read to produce one aggregated measure of electrical use per household or business address every two months, the meters available on the market today allow the option of using sensors to disaggregate overall electricity consumption in order to discern

---

71. Interview, Seattle City Light Personnel, Seattle, Wash. (Mar. 29, 2015).

72. *Id.*

73. *Id.* The Pilot Project Summary and Conclusions are on file with authors.

74. *Id.*

75. *Seattle City Light First Utility Certified for Green Button Data*, SEATTLE.GOV, <http://powerlines.seattle.gov/2014/06/20/seattle-city-light-first-utility-certified-for-green-button-data>.

the use of identifiable electronic appliances.<sup>76</sup> This type of sensor gives users and utilities the option of viewing the consequences of appliance use in terms of electrical demand in real-time.

Beyond allowing users to respond to and manage demand, Seattle City Light personnel also describe the potential benefits of this new technology in terms of the ability to more precisely discern where electricity is flowing, to re-route electricity based on this information, to improve the management of voltage issues and problems in the system, and to ensure a smooth flow of electricity.<sup>77</sup> This will also allow the utility to identify more precisely where in the system people may be tapping electricity illegally. Of course, as all of this data becomes more detailed, reporting electrical consumption over time or by appliance, it carries a greater potential to compromise the privacy and security of the home and workplace.

e) Department of Transportation

Transportation assets are expensive to build, operate, and maintain, and until recently, transportation departments have also had to spend inordinate amounts of money, time, and labor to simply collect data to estimate how much we use the various components of our transportation networks. The integration of GPS technology in smart devices on our person or in our cars has fundamentally transformed this problem for the Department of Transportation from one of costly and time-consuming data collection, to one of concern about the privacy implications of collecting and using personalized data. For example, the City has contracted the services of Parkeon to operate pay stations that accept credit card payments for parking,<sup>78</sup> and has recently added the services of Pay by Phone, a mobile payment vendor. In these cases the vendors develop databases that contain vehicle information and the identities of parking permit purchasers. The vendor attempts to anonymize the data by removing a subset of fields, and feeds the resulting dataset back to the department.

In regard to travel behavior, we found two opposing approaches to data collection underway in the Department. One unit within the Department contracts with the fitness software company Strava to provide data describing the movements of individuals who have opted in to the use of their running and cycling app.<sup>79</sup> Another unit in the Department has been using, through the vendor Acyclica, Bluetooth and Wi-Fi readers installed

---

76. For an explanation of this process, see UWTV, *UW Four Peaks -- Shwetak Patel*, YOUTUBE, <https://www.youtube.com/watch?v=nnzzTFs0O2g>.

77. Telephone Interview, Seattle City Light Personnel (Apr. 7, 2015).

78. *Id.*

79. *Id.*

in public places that automatically read and record the Media Access Control (MAC) addresses of multiple devices—i.e., smartphones, laptops, and automobile computers—to track the movement of individuals across the city. A MAC address is a serial number assigned to a computing device, typically during the manufacturing process, to make that device uniquely identifiable from all other network devices in the world. When turned on, personal computing devices constantly send their MAC addresses in signals that perform an electronic handshake with Bluetooth and Wi-Fi routers. In this case, Acyclica has been granted permission from the City to install readers that “sniff” and send the unique MAC identifier of personal devices to the servers of the firm. The firm, in turn, sends the data it collects on personal travel behavior to the Department.<sup>80</sup> Though people have no obvious way of knowing that their movements are tracked by Acyclica’s devices, the firm operates a web-based portal that allows anyone with a MAC address to retrieve the travel behavior data specific to that device.<sup>81</sup>

f) Police Department

With respect to open data, the Seattle Police Department is a self-described “manufacturer of data for the public.”<sup>82</sup> In terms of the demand for data, people have always expressed an interest in police activities, listening to police scanners, and requesting incident reports and data from 911 calls. The Department has adopted multiple technologies with implications for the generation of big data: they have a cloud-based service that captures citizens’ online reporting, they deploy smart phones, they have computers onboard vehicles, they generate in-car video and body camera video, and are proposing to develop a data analytics platform with multiple applications. The Police Department typically receives three times more public record disclosure requests than any other department in the city.<sup>83</sup> In the first quarter of 2015, the number of requests rose by 400%, to an estimated 2,500.<sup>84</sup>

Personnel in the Seattle Police Department note that the rising increase in demand for public disclosure has coincided with the digitization of files and the advent of video recording devices mounted on the dashboards of

---

80. Interview, Seattle Department of Transportation Personnel, Seattle, Wash. (Mar. 10, 2015).

81. See *Analyzer User Guide*, ACYCLICA, <https://acyclica.com/support/documentation> (last visited March 12, 2015). The web portal is available at <https://acyclica.com/products/acyclica-analyzer>.

82. Interview, Seattle Police Department Personnel, Seattle, Wash. (Jan. 14, 2015).

83. Interview, Seattle Police Department Personnel, Seattle, Wash. (Mar. 5, 2015).

84. *Id.*

patrol cars and worn on the bodies of officers.<sup>85</sup> Gradual shifts over time have allowed public disclosure requests to become anonymous and free of charge. Individuals in the department explained that people making public disclosure requests used to have to provide a phone number to call, so that people would be notified when the documents were ready, or could be called to clarify the request. As one interviewee explained:

[T]he department has moved from paper to electronic, so the people think it should be accessible data, they think that a report should be available right away, even though there are protocols. The types of records [now include] body cams, in-car video, 911 calls, audio statements in the field, photos, officers receive video, text messaging, emails, web browsing. People expect to be able to access this information as much as they want in real time.<sup>86</sup>

Departmental personnel explained how demands rise “on the back end” with the number of public disclosure requests.<sup>87</sup> The department receives about 125 requests per week. The department employs seven people full-time to respond to public disclosure requests, plus additional attorneys, paralegals, and people dedicated to 911 and video requests.<sup>88</sup> Each request to the department generates a series of actions and corresponding logs. Detectives assigned to the relevant case participate in the process, helping review requested information for civilian safety, privacy, officer safety and for compliance with numerous other policies and regulations that pertain to police records. The personnel involved are “very careful and conscious of the fact that we are dealing with victims and the most vulnerable and not on their best day.” As they explain, “we want victims to continue to cooperate with the department, all weighting this with trying to be as open as we can.” People are given the data they have the authority to receive (e.g., victims receive different data than the media). When data is not released, officers are required to explain the reasons in an exemption log.

The Police Department is struggling with the demands created by the sheer volume of both footage and requests. Specifically, the Department must wrestle with privacy concerns stemming from the fact that body camera video contains recordings of persons other than the police officer. In its most recent move, as part of the recently initiated program using body worn video cameras, the department has launched its own YouTube

---

85. *Id.*

86. *Id.*

87. Work “on the back end” consists of the tasks that Department personnel must carry out in order to satisfy a public records request.

88. *Id.*

channel.<sup>89</sup> Besides posting raw video clips that have been processed for public disclosure, the department is blurring video content and deleting audio (to “redact” the identity of persons in the video) that has not been through the process, and posting these feeds to YouTube to facilitate public disclosure, with dates, times and incident numbers so that interested parties can see what is available and make more specific requests.

g) Parks and Recreation

Seattle Parks and Recreation maintains twenty-six community centers and organizes hundreds of volunteers to provide community services and events. Those events are attended by thousands of children and adults registered in their databases each year.<sup>90</sup> The Department takes a conservative approach to public disclosure requests. Personnel have been successful in redacting the information describing the people who volunteer to run and attend their programs and, under the law, the City has the discretion to redact considerable amounts of information pertaining to juveniles.

Perhaps as a result of working predominantly with youth and at-risk populations, such as special needs children, Department personnel expressed the need to be careful when releasing information for public consumption.<sup>91</sup> The Department has sensitive information about employees, volunteers, and adults and youth registered for programs. They are aware that the use of personal information, when distributed through either open data portals or in response to public disclosure requests, can give people the information they would need to be able to harass someone, stalk someone, seek revenge, and commit various crimes. Personnel described fights between individuals, a person stalking a volunteer, and community groups pitted against one another over a controversial park project, as examples of circumstances that have precipitated public disclosure requests for personal information. Personnel described their success disclosing incident reports to requestors, while redacting the information that could be used to contact the other party.

h) Fire Department

The Seattle Fire Department manages large amounts of data, but has not yet gravitated to new information technologies to the degree that Seattle City Light, the Department of Transportation, and the Police Department

---

89. *SPD BodyWornVideo*, YOUTUBE, <https://www.youtube.com/channel/UCcdSPRNt1HmzkTL9aSDfKuA> (last visited July 22, 2015).

90. Interview, Seattle Parks and Recreation Personnel, Seattle, Wash. (Mar. 5, 2015).

91. *Id.*

have. Unlike other departments, the Fire Department provides emergency medical services, and controls the release of medical data in accordance with The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and related rules and regulations governing personal health information.

Approximately 80–90% of Fire Department responses to calls are medically related.<sup>92</sup> In these situations Fire Department personnel produce paper and carbon copy medical reports that they input into special HIPAA-compliant scanning devices. About 200 two-page medical reports have to be entered each day.<sup>93</sup> Before it is stored in Department databases, data is shared and reviewed by the Department, the station that responded to the call, and with University of Washington doctors working with the Fire Department. It reportedly takes about ninety days before these records enter the Department databases. Department personnel suspect that the movement to digitize this process is not likely to change the demand for public disclosure of these records because requestors have to provide proof of identification, such as a scanned copy of a driver's license, to receive a copy of a report.

The Fire Department also stores sensitive data that does not pertain to HIPAA. And, like other departments, it receives requests that appear “frivolous.”<sup>94</sup> Interviewees explained that a person could make a targeted use of the law to inundate the Department with requests. Even though a request appears frivolous, “you are legally required to respond . . . but we can't possibly respond.” The PRA requires a response within five days of every request. The fine for missing this window, can reach as much \$100 per page, per day. “It's the only hard deadline and [someone] could try to get you to trip up and you have to hit respond to those. Some of them could be months' worth of work. [Someone could] then send a message to the council threatening to sue and say you are not in compliance with the PRA.”<sup>95</sup>

The Fire Department, like other departments, is experiencing pressure to release data in the form of public disclosure requests. However, accustomed to maintaining medical and other sensitive data on paper and specialized electronic systems, this department realizes many such requests may not be justifiable.

### 3. *Analysis*

The one common approach departments have in regards to open data is the desire to reduce the financial cost of public disclosure. If pushing data

---

92. Telephone Interview, Seattle Fire Department Personnel (Mar. 19, 2015).

93. *Id.*

94. *Id.*

95. *Id.*

to the open data portal, a YouTube channel, or a more sophisticated portal such as the Green Button initiative, promises to reduce the cost of responding to public disclosure requests, then departments generally aim to do so.

Departments differ widely, however, in their pace and degree of adoption of smart technologies, and thus they differ in terms of the challenges they face in preserving privacy and social justice when data is pulled for public disclosure from city files. Departmental personnel appeared interested in serving the public interest and fostering transparency. Many also share concerns that the PRA can be, or perhaps already is being used for, self-interested, wasteful, or harmful purposes. The timing of the growth of such requests coincides with the transition from paper to digital records, from charging a nominal fee to copy records to providing them at no charge, and from named to anonymous requests. The piecemeal exemptions to public disclosure that have accrued in the PRA show that some departments have tried to solve the problem through the State Legislature. The PRA includes a lengthy list of data exempt from release, categorizing exemptions based on specifically named attributes (e.g., name, address, telephone number) in the data, subjects represented by the data, and public programs or other contexts that motivated the public collection and disclosure of the data.<sup>96</sup> Other departments have taken a slower

---

96. The Public Records Act lists types of data exempt from public disclosure and, in doing so, either names specific attributes or uses the broader term “personally identifying information” to specify the data that are to be exempt. For example, in a section pertaining to public utilities and transportation information, exemptions include:

addresses, telephone numbers, electronic contact information, and customer-specific utility usage and billing information in increments less than a billing cycle of the customers of a public utility contained in the records or lists held by the public utility of which they are customers, except that this information may be released to the division of child support or the agency or firm providing child support enforcement for another state under Title IV-D of the federal social security act, for the establishment, enforcement, or modification of a support order.

WASH. REV. CODE § 42.56.330(2) (2014). Further on, in the same section, exemptions include:

The personally identifying information of persons who acquire and use transponders or other technology to facilitate payment of tolls. This information may be disclosed in aggregate form as long as the data does not contain any personally identifying information. For these purposes aggregate data may include the census tract of the account holder as long as any individual personally identifying information is not released. Personally identifying information may be released to law enforcement agencies only for toll enforcement purposes. Personally identifying

approach to adopting technology, concerned about the very same implications. A few have been more deliberative in their service of public disclosure requests, taking a more proactive stance of exempting personal information from public disclosure requests.

Interviewees' conceptions of the market for municipal data varied. When favoring the commercial application of open data, interviewees' conceptions of the firm appeared to be aligned with small startups and newly created firms. The idea of pushing data to an open platform for commercial use is not universally embraced, however. Many interviewees questioned the idea that it is possible to favor the interests of some firms, such as small startups, over others, when data made open is open to all. Those concerned with the differential treatment of firms seemed to have a broader view of the market for municipal data, including large, well-apportioned organizations. Only the Police Department expressed awareness of the way data brokers use publicly disclosed data—an issue raised because of the uses of profiles in criminal investigations. Contractual relationships between the city and firms cloud these issues. Finance and Administrative Services, for example, raised the issue of the unintended spilling of data by the vendors under contract to the city to create online data portals. Seattle City Light will face the same issues in designing portals for advanced metering data.

### C. TECHNICAL ASSESSMENT: OPEN DATA ANALYSIS

This section explains the technical analyses we conducted on the City of Seattle's current municipal open data. At issue is the question of how the city may evaluate, prior to release, the potential for a dataset to compromise privacy.

#### 1. *The Problem of Cumulative Risk of Re-Identification*

From our initial interviews we learned that most datasets released by the City of Seattle on the open data portal had received some scrutiny with regard to potential privacy harms. However, the practices in place only modeled the risk of data releases for each dataset in isolation.

As various scholars have found, otherwise innocuous datasets can be joined together in ways that result in re-identification and breaches of privacy. This simple fact, evidenced by the accomplishments and practices of firms that have amassed detailed dossiers on millions of people, is reason

---

information may be released to law enforcement agencies for other purposes only if the request is accompanied by a court order.

§ 42.56.330(7).

to question the ability of a municipality to release any one dataset about persons while preserving the anonymity of those persons.<sup>97</sup>

Public policy reflects the idea that the potential harm caused by releases of personal information is a function of what the combination of two or more pieces of information may reveal about an individual. This is expressed in various state laws by the way in which they approach Personally Identifiable Information (PII),<sup>98</sup> typically defined as the combination of two or more attributes for the purpose of protecting individuals' privacy, identity and personal safety.<sup>99</sup> The City's policies and regulatory framework for governing the release of data generally follow this line of reasoning. As illustrated by its release of business license data, the City of Seattle correctly and appropriately uses this criterion to manage the issue of potential privacy harm in their analysis of each dataset prior to publication. However, this is an analysis of a dataset in isolation.

The fact that multiple datasets can potentially be joined together using matching information in common fields threatens the validity of any risk assessment that has been limited to a single set of data. All that an actor would have to do to invalidate the claim that the release of any one dataset is risk-free is to join it across common fields with identical or similar data.

---

97. See Ohm, *supra* note 11; Narayanan & Shmatikov, *supra* note 11, at 24–26; Solon Barocas & Helen Nissenbaum, *Big Data's End Run around Anonymity and Consent*, in *PRIVACY BIG DATA, AND THE PUBLIC GOOD*, 44–75 (Julia Lane et al. eds., 2014).

98. *Security Breach Notification Chart*, PERKINS COIE, <https://www.perkinscoie.com/en/news-insights/security-breach-notification-chart.html> (last visited July 21, 2015) (providing a full list of state definitions of PII, current as of June 2015).

99. See NIST, *GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII)*, <http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf>. The NIST Guide defines PII to include:

[A]ny information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information.

*Id.* See also Narayanan & Schmatikov, *supra* note 11, at 24. Narayanan and Schmatikov note:

PII is surprisingly difficult to define. One legal context is provided by breach-notification laws. California Senate Bill 1386 is a representative example: its definition of personal information includes Social Security numbers, driver's license numbers, financial accounts, but not, for example, email addresses or telephone numbers. These laws were enacted in response to security breaches involving customer data that could enable identity theft.

*Id.*

The resulting merged dataset would not have to be a successful join of every record in order to be used to re-identify individuals, or to associate persons with attributes that threaten to compromise privacy or safety. In other words, cities looking to release public data responsibly face the need to develop their capacity to assess the privacy posture of collections of datasets more globally, encompassing the impact that additional releases may have in combination with existing corpuses of publicly, and perhaps privately available data.

## 2. *A Proposed Method of Ex Ante Evaluation*

Our research includes an analysis of the tabular data already released and publicly available at Seattle.gov. The research design presented here models the methods that could be used to assess the privacy of collections of datasets before they are released from municipalities.<sup>100</sup>

Someone wishing to identify potential privacy-violating joins must first take the step of identifying what joins are possible. Traditional database joins involve simply combining records from one table with another based on a known shared field. Our aim, however, is to discern the maximum possible extent of joins. So, in contrast to traditional approaches, the joins we are contemplating combine information, which may not be perfectly matched, or may be nominally classified as different. The purpose is to produce the greatest possible degree of connections across datasets that have been published separately. For example, fields with differing data types, or combinations of fields such as latitude and longitude can be joined across datasets with a field called “address” if sufficiently overlapping information is compared.

A second step is to then assess identified joins for their potential harms to privacy. To accomplish this, some care must be taken to correctly categorize and classify the types of information in the datasets. The analysis depends on an understanding of the harms made possible through the association of different attributes, as they are found in the published datasets and joined using the methods described above. Rules and regulations governing personally identifiable information offer limited guidance;<sup>101</sup>

---

100. Anyone in the City interested in evaluating an additional dataset prior to release would add that dataset to the corpus of existing public data and repeat the analysis. It is important to note, however, that our analysis was limited in time and resources. It represents a starting point for further research.

101. See Narayanan & Schmatikov, *supra* note 11, at 25 (“What is ‘reasonable’? This is left open to interpretation by case law. We are not aware of any court decisions that define identifiability in the context of HIPAA.”).

empirical cases of re-identification are more likely to inform this part of the exercise.

These two steps are encapsulated in Rob Kitchin's definitions of indexical and attribute data. Indexical data is important because it enables attributes to be linked, and often is the data that can be used to identify the subject of the attribute.<sup>102</sup> Unique identifiers such as passport numbers, account numbers, MAC addresses, order and shipping numbers, and manufacturing serial numbers are examples of indexical data, as well as names, addresses, and zip codes. What people and firms are joining together with the use of indexical data are attributes that describe the subjects of the data. As Kitchin notes, "Attribute data are data that represent aspects of a phenomenon, but are not indexical in nature. For example, with respect to a person the indexical data might be a fingerprint or DNA sequence, with associated attribute data being age, sex, height, weight, eye colour, blood group, and so on."<sup>103</sup> The vast bulk of data in storage are attribute data, and because the attributes that may be sensitive in terms of privacy or social justice are associated with various indexical fields, this association places sensitive data at risk.

The expansion of indexical fields gives rise to new and more expansive datasets, along with rising hazards to privacy and social justice. In addition to these factors, the adoption of advanced technologies further thickens the flow of information, with more opportunity to join or enrich existing datasets with potentially compromising information. Kitchin mentions how the ingenuity and economic drive of people and firms to find more and more ways to join data has resulted in the expansion of fields considered useful for indexing.<sup>104</sup> Thus the threat of re-identification with the release of data is a moving target. As more variables become useful for indexing, more publicly available datasets may be used to join datasets in previously unimagined ways.

One way to operationalize the first step—determining which joins are possible—is to turn collections of tabular datasets into network graphs that illustrate a variety of strategies for identifying potential joins between multiple datasets. This approach casts individual tables (i.e., each a dataset) as nodes in a network, connected by lines as identified by a specific join identification strategy (e.g., joining tables on the basis of specific indexical fields, such as location in space, as identified through latitude and longitude). If each separate table were joined on one indexical variable,

---

102. See ROB KITCHIN, *THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES* 8 (2014).

103. *Id.*

104. *Id.*

showing tables as nodes and indexical field data on the lines connecting nodes to one another, one could see within the scope of a single diagram the possibility for joining multiple datasets. With a diagram showing the potential to join multiple datasets along one or more indexical fields, determining the possibility of connecting an attribute in one table to an attribute in another table could then become a network pathfinding operation. The network of datasets resulting from this approach would be amenable to the full-range of network analytical methods.<sup>105</sup> New datasets under consideration for release could be added to the network, and the changes in network topology studied with precision.

The second step—the assessment of the potential for harm from any one specific join—is likely to remain somewhat of a human intelligence task. This approach segments individual attributes into a continuum of privacy and social justice risk. Combining this continuum with a network dataset could allow the programmatic identification of instances where connections between low-risk attributes (e.g., describing the built environment) and high-risk attributes (e.g., describing persons in the built environment) result in potential information leaks.

### 3. *Potential Join Strategies*

We have envisaged several join identification strategies, all of which have different characteristics, advantages, and disadvantages with respect to quality of results, false positive or negative rates, processing time, and computing resources.

Some of these strategies work at the schema level (i.e., across field names or column headings, in the case of tabular data), and compare the names of individual fields (e.g., latitude, longitude, address). These strategies may be especially useful for inferring links between datasets that are held by a city and datasets that may not be wholly obtainable by a city (i.e., held by a third party). For example, one could infer a potential join where two tables share an “address” column. Other strategies extend the schema comparison approach by using natural language processing to identify conceptually related terms, inferring matches between fields such as “location” and “postal address.”

Other strategies that are more exhaustive operate at the level of the data itself. These include the attempt to join, through exact matching, all fields in all datasets. This is computationally expensive, but answers concretely the question of where deterministic joins are possible. Other variants of this

---

105. An example of an analytical method that could be applied is Dijkstra's shortest path algorithm. See E. W. Dijkstra, *A Note on Two Problems in Connexion with Graphs*, 1 NUMERISCHE MATHEMATIK 269 (1959).

strategy include spatial joins, for example, that make geometric comparisons of the spatial attributes within tables.

Many more join identification strategies are likely to be employed by data brokers, or other would-be users of these datasets. Future work might identify additional strategies or integrate ensembles of strategies for identifying potential joins, such as using natural language processing techniques to perform meaning-based comparisons of all fields in all databases.

#### 4. *Analysis and Results*

We implemented several join identification strategies, and used them to perform an initial analysis of the datasets that were publicly available from the City of Seattle's open data portal, as of April 1, 2015. At that time, there were 235 datasets on the Socrata open data portal from the City of Seattle. The strategies we employed include:

- Exact match of field name
- Tokenized match of field name components<sup>106</sup>
- Levenshtein distance match of field name<sup>107</sup>
- Natural language processing match of field name (i.e., Wordnet)<sup>108</sup>
- Exhaustive exact match of column contents

---

106. Technopedia offers the following definition of “Tokenization”:

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining.

*Tokenization*, TECHOPEDIA, <http://www.techopedia.com/definition/13698/tokenization> (last visited July 23, 2015).

107. The Levenshtein Distance can be defined as “[t]he smallest number of insertions, deletions, and substitutions required to change one string or tree into another.” *Levenshtein Distance*, NIST, <https://xlinux.nist.gov/dads/HTML/Levenshtein.html> (last updated June 22, 2015); *see also* *Levenshtein*, PHP MANUAL, <http://php.net/manual/en/function levenshtein.php> (last visited July 23, 2015) (“The Levenshtein distance is defined as the minimal number of characters you have to replace, insert or delete to transform str1 into str2.”).

108. *The Stanford Wordnet Project*, <http://ai.stanford.edu/~rion/swn/> (last accessed July 23, 2015) (“By applying a learning algorithm to parsed text, we have developed methods that can automatically identify the concepts in the text and the relations between them.”); *see also* Snow et al., *Learning Syntactic Patterns for Automatic Hypernym Discovery* (2004 Conference on Advances in Neural Information Processing Systems), [http://ai.stanford.edu/~rion/papers/hypernym\\_nips05.pdf](http://ai.stanford.edu/~rion/papers/hypernym_nips05.pdf).

- Partial latitude and longitude geometric match of geospatial column contents

Relatedly, we have partial results of an ordering of the individual fields found within Seattle's open datasets. The number of datasets with tabular data that could be analyzed (i.e., contained field names and field contents) was 204. The City offices contributing to the corpus of open data included: City Budget Office; Department of Human Services; Department of Neighborhoods; Department of Planning and Development; Seattle Fire Department; Office of the City Clerk; Seattle Police Department; Office of the Mayor; Seattle City Attorney's Office; Department of Information Technology; Department of Transportation; Finance and Administrative Services; Seattle Public Utilities; and the Seattle City Council.

The datasets contained a wide variety of information, such as building permits, electrical permits, land use permits, code violations, surveys of residents' use of information technology, traffic counts, announcements of learning programs and events, commute trip reduction surveys, police department incident reports, active business licenses, 911 call logs, housing emergency responses, logs of police in-car video, grants and funding, adopted budgets, and neighborhood matching grant reports. Many were inventories of infrastructure assets, such as assets listed for auction, cultural spaces, road weather information systems, trails, street parking signs, and neighborhood maps. Of note are several datasets on the Socrata portal that are produced as part of a performance dashboard for municipal services.<sup>109</sup> Performance dashboard datasets include, for example, pothole complaints and repairs, streetlights data, conservation data, planted trees, first arriving engines in emergency response, police reported collisions, bus ridership, city building energy use data, pea-patch garden registrants, residential burglaries, motor vehicle theft, and civil rights performance data.

#### a) Joins Using Exact and Flexible Matching Strategies

As one would expect, exact matching strategies (i.e., exact matches of field names, or column headings) for these datasets appear to result in many false-negatives, whereas more flexible matching strategies appear to result in many more false-positives. For the purpose of demonstrating potential flaws in vetting datasets for publication, flexible strategies are important to use so as to not overlook valid matches; eliminating false positives manually was the price for complete coverage.

Results from our schema-based join identification strategies suggest a great deal of connectivity between datasets on Seattle's open data portal.

---

109. To explore these datasets, and others, see *Performance Seattle*, SEATTLE.GOV, <https://performance.seattle.gov>.

The total number of field names in the corpus of 204 datasets was 3,859, and the number of unique field names (a product of exact match of field name) was 1,981. Tokenized, the field names in the corpus of datasets produced 6,061 parsed names. Among these were many duplicates. Eliminating duplicates left 1,828 parsed field names. The Wordnet comparison of parsed field names returned thirty-one pairs with 100% match, and another 230 pairs with a 50% match.<sup>110</sup> For example, forty-six fields are named “address.” Given the ubiquity of certain terms such as address, as well as other common fields, the number of connectable tables results in a network graph that expresses the possibility of joining nearly all tables in the set—forming one comprehensive table out of 204. This validates the premise that it is possible to recombine data in ways that violate the current model for vetting publication of datasets (i.e., assessing datasets in isolation).

Results from our content-based join identification strategies were also promising. We performed a many-to-many comparison (i.e., an exhaustive comparison of data entries in all cells), using exact matches only, across all fields of all datasets. This resulted in a large number of irrelevant matches for common objects (e.g., numbers, “true/false,” “yes/no”), and very few exact matches for data in cells. This result was expected, since the published datasets do not constrain or normalize data in fields. For example, reliance on exact matches produces results that suggest “302 N Baker Street” is not an exact match to “302 N Baker St.” This supports the notion that using broader, more flexible strategies for finding matches and weeding out false positives is a useful approach.

After the exhaustive join on exact matches of field contents, the next likely research step was to either use more flexible joining strategies with the entire corpus of data, or more targeted joins on the basis of potential privacy harm. We opted to implement the latter, through one smaller but significant strategy for joins, with the purpose of illustrating some of the unusual qualities of local government data.

#### b) The Special Relationship Between Municipalities and Spatial Data

The more we studied the open datasets, the more it appeared to us that spatial data is highly represented among Seattle’s municipal open datasets. We mentioned the commonality of “address” but it is worth noting that nearly all of the datasets included spatial data of one kind or another (i.e.,

---

110. Results show how closely Wordnet’s system believes they are related to one another. The parsed field names included in this analysis were all nouns. All other parts of speech were excluded.

latitude, longitude, block, location, mailing, shape, zip code, acres, area, and shape files).

There is a logical rationale for this observation. If, as employees of departments had suggested in interviews, efforts to de-identify datasets prior to publication primarily involved the removal of names, telephone numbers, and email addresses, while retaining street address (sometimes aggregated to the nearest 100 block), zip code, or another similar spatial identifier, then spatial data would be more likely to be retained in the datasets made public. Also, considering that cities are primarily interested in data regarding activities within the spatial boundaries of their jurisdiction, and meaningful determinations of demand, supply, and quality of services often pertain to the delivery of services across the spatial extent of the jurisdiction, spatial data is likely to be a key variable in municipal data.

However, spatial data can also be the means to identify individual home and business owners and occupants. Residents are readily identified when their name is associated in any publicly available dataset with these properties. For example, the City of Seattle includes the names of persons on building permit applications in open datasets, and King County (which includes the spatial extent of Seattle) maintains a publicly available dataset that includes the names of the owners, addresses, and assessed value of the properties.

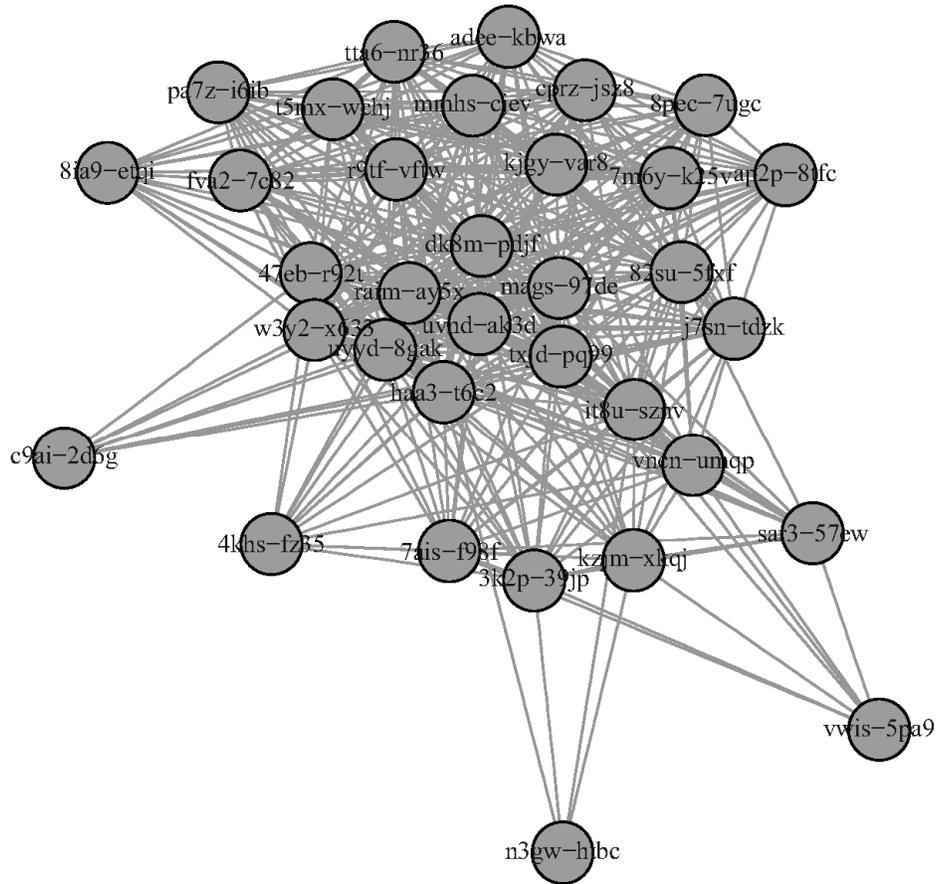


Figure 1: Results of 5-Meter Spatial Join of Latitude and Longitude Column Contents.<sup>111</sup>

On this basis, we conducted a simple spatial join of datasets sharing the field names of latitude and longitude. For this procedure, we drew a circle, 5 meters in diameter around each point in space identified in columns with the heading latitude and longitude (both of which were present in 34 of the 204 tabular datasets available). If the point from one dataset was found within the circle of a point from another dataset, this constituted a join between the two datasets.<sup>112</sup> Joins between two datasets, measured in this

111. Datasets in Figure 1 are represented by circles with alphanumeric identifiers. Datasets are linked to one another in the network graph when six or more location matches, in a 5-meter radius of one another, occur between the datasets. Data collected from all Tabular datasets on the City of Seattle's Open Data Portal, as of April 1, 2015.

112. Analysis was carried out using PostGIS, with an overall program logic instrumented in a combo of Python and Bash. Overview of steps in the analysis:

way, are highly likely to be referring to the same parcel or piece of property. The results are shown in Figure 1.

In the figure, nodes correspond to datasets, and are labeled with the alphanumeric identifiers of datasets used on the Socrata platform. The lines connecting the nodes indicate matches between datasets. Links were removed when the number of matches was less than six, thus all lines indicate more than six matches between datasets. From this visualization one can assume that nearly all tables in this sample of tables ( $n = 33$ ) will have spatial matches.

The meaningfulness of the match depends on the context of the locations matched. Manual inspection of field names and titles of the sample datasets suggests that the spatial locations matched are perhaps public facilities (e.g., community centers hosting multiple types of events, locations of sensors for data collection such as bicycle and other traffic counts) but also private facilities (e.g., locations undergoing repeated building inspections and permitting procedures, locations identified in multiple events such as 911 calls for police and fire). In this research agenda, the next step would be to conduct more flexible comparisons where, for example, latitude and longitude are geocoded and compared to street addresses or other forms of location information.

#### c) Attributes on a Continuum of Personalization

In terms of the potential for privacy harm, a very limited scan of attributes amongst datasets, both within and outside municipal open data for Seattle, produced a rather rich set of information for the purpose of profiling individuals. Limited only to three datasets in Seattle and a fourth in King County, these attributes suggest how weaknesses in the ability to

- 
1. Convert lat/lon text strings into WGS84 Geometries (a reference datum used by Socrata).
  2. Create new empty geometry field.
  3. Translate points into NAD83(HARN) Washington State Plane N format, meter units.
  4. Create 5 meter buffer around points. This value was chosen somewhat arbitrarily to allow matches of points that differ only by the floating point precision of the lat/lon. This distance was generous enough to smooth over any minor discrepancies in parcel size, but conservative enough that any identified matches would pretty much be a stones throw from each other.
  5. Construct spatial indexes using GiST strategy.
  6. Identify matches based on the condition of intersection between any two circular buffers (ST\_Intersects function).
  7. Return count of matches.

effectively de-identify individuals through the elimination of indexical fields and the aggregation of data across space could result in serious consequences in terms of privacy and social equity.

Table 1: Attributes from Four Open Data Sets on a Continuum of Personalization.<sup>113</sup>

Fields	Datasets				Potential Privacy Concern
	Property Value	Tech. User Survey	Business License	Building Permits	
Name	●		●	●	Persons
Address/Location	●			●	
Phone Number			●		
Age		●			
Gender		●			
Income		●			
Home Value	●				
Zip Code		●			
Sexual Orientation		●			
Race		●			
Level of Education		●			
Language		●			
Number in Household		●			
Employment		●			
Unpermitted Activity				●	
Internet Use		●			
Uses of Cable		●			
Incident Type/Descrip.			●		
Permitted Activity				●	
Value of Alteration				●	
Permit Type				●	

The City of Seattle datasets represented in Table 1 include permitting data from the Department of Planning and Development, Business License Data from Financial and Administrative Services, and the Department of Information Technology's survey of resident uses of information technology ( $n = 2900$  residents surveyed). King County's public dataset showing property ownership and tax assessment is also included. Note the ability to

113. Some fields of Table 1 contain data that may be used to identify persons or infer the identity of persons. Some fields contain data that may be used to categorize persons into racial, social, or economic groups. Government data contains many additional fields of data with as yet unknown implications for privacy. Fields that may be used to identify (e.g., name, address) or infer the identity of persons (e.g., age, gender, zip code) are indexical, and can be used to join these data into one universal set to form dossiers on individuals, groups of individuals, households, and neighborhoods.

join the property value, business license, and permitting databases using the names of the property and business owners. This one act brings together name and contact information, such as address and phone number.<sup>114</sup> While there is no obvious overlap of fields between the technology user survey, and other datasets, it is worth noting that one of the more popular and widely used indexical fields for re-identification is zip code. With the plethora of demographic fields provided in the survey dataset, it is not difficult to imagine a data broker or similar type of firm using zip code to join and re-identify survey respondents. At the very least, the privacy implicating and highly differentiated fields in the survey could make this dataset a desirable target for commercial interests seeking to re-identify subjects and enrich their existing dossiers on city residents.

#### d) One Simple Example of a Profile

Finally, to demonstrate the kind of personal profile which can be gathered today from open data published by the City of Seattle, we chose a single location and produced joins from eight Seattle open datasets. The information gathered from these datasets revealed:

1. Property owner's full name (multiple spellings)
2. Multiple major building projects, most with associated code violations related to follow-up and/or inspections
3. Junk storage violations
4. Vacant building-related issues
5. A fire in the main structure

There is enough information in any one of these datasets to join this profile with the King County dataset that shows the assessed value of property, which may be used as a proxy for wealth or income. The property is among those in the city that have received the lowest possible valuation.

There is distress involved in some of the revealed incidents as well as loss of personal property and net worth, all tied to dates, times and a specific person's name. The level of information revealed from the combination of these eight open data sets—all indexed using spatial location—is more than most individuals would be comfortable with.

#### 5. *Open Data Assessment in Sum*

These technical assessments suggest the extent to which the release of multiple, seemingly benign municipal open datasets holds the potential to compromise privacy, or pose threats to social justice. The City of Seattle,

---

114. This emphasizes the importance of excluding licenses for businesses located in residences from open data.

however, like many cities in the U.S., governs many more datasets than those currently available as open data. Many of those datasets are produced, processed, copied, and stored in the information systems of firms under contract with the City.

#### D. LEGAL ASSESSMENT: VENDOR CONTRACTS

The preceding section describes risk as a function of technical processes, demonstrating how data that is “safe” in isolation may yield more private details than anticipated when combined or correlated. In this section, we describe risk of another sort: the risk associated with turning over the processing and storage of resident data to third party vendors. Cities use vendors extensively. And vendors have different capabilities and incentives than a municipal government; they may be more or less capable of keeping data secure, and are not likely to be as responsive to residents as their city government. As our qualitative analysis makes clear, stakeholders will ultimately hold cities responsible as custodians and expect them to uphold constituent values.

The relationship between the City of Seattle and its vendors is described in its contracts. We therefore undertook an analysis of a carefully selected sampling of contracts between the City and its vendors. The goal of this research was to determine whether vendors with access to City data—including data about employees and citizens—were contractually obligated to engage in best practices around privacy and security, thus preventing the unintended spilling of data. We found that some were, and others were not. This does not necessarily mean that any vendor engages in bad behavior, only that they do not make commitments that help foreclose the possibility. On the basis of this work, we later recommend that the City generate a standard contract including privacy and security language to use as a starting point for any future outsourcing of data processing, gathering, or storage.

Among the insights we gleaned from our focus group sessions were that residents did not tend to differentiate between the specific constructs of open government or public records requests and the city’s role in general as a custodian of resident data. The city collects, stores, processes, and in some instances shares information. Although we have developed a taxonomy of push, pull, and spill in this paper, the picture for residents seems rather less differentiated.

In general, we found that relatively few vendor contracts made guarantees around the privacy or security of resident or employee data, and that the contracts that did make such guarantees did not use anything like the same language. There was no “smoking gun,” in the form of a highly irresponsible provision, but there were places where due diligence might

have recommended changes to allay stakeholder fears and concerns. The findings that follow form the basis of our recommendation, *infra* Section IV.G, that the City develop a standard vendor agreement that incorporates baseline or default provisions regarding how information is accessed, shared, and secured.

Residents want to feel as though cities are using information wisely to their benefit across the board. Cities do not collect, process, or store information on their own. Like all major enterprises, they work with partners. Accordingly, the circle of trust regarding municipal data is wider than just a city itself—it includes their providers. Cities entrust resident data to providers for a variety of purposes, including storage, analysis, and connectivity. For example, the City of Seattle Police Department works with Evidence.com—a subsidiary of Taser—to store video from police lapel cameras. Seattle employees work with Verizon and Motorola to communicate. As noted previously, the City’s existing open data portal is managed by Socrata.

The primary means by which cities can maintain its trust with residents in light of these partnerships is by getting these providers to agree to a comparable level of responsibility and data hygiene. Indeed, the city’s relationships with vendors are governed by terms of service, privacy policies, and other service agreements.

We undertook to examine these documents in an effort to assess whether they respect privacy and security by their terms. Our method involved selecting eighteen particularly important master agreements (plus sub-documentation) from five departments. We based this selection on the in-depth interviews we conducted with employees across the City. An attorney in private practice analyzed the documents according to parameters set by a member of our team with deep experience in privacy law, specifically including privacy policies and terms of service. That team member then reviewed and synthesized the findings for presentation here.

### 1. *Privacy*

We first looked for language addressing what if any rights the subjects of data being processed by the City’s partners may have. In the consumer privacy context, such rights generally include understanding what information has been collected and why, how it is secured, with whom it is shared, and so on. A good benchmark is the set of obligations imposed on websites under California’s privacy notice law.<sup>115</sup>

---

115. See CAL. BUS. & PROF. CODE § 22575 (West 2014). See also CALIFORNIA ATTORNEY GENERAL, MAKING YOUR PRIVACY PRACTICES PUBLIC:

The picture on privacy was mixed. Whereas some providers specifically reference the ability of data subjects to access their data (e.g., Paybyphone, Volgistics, and Microsoft), many others made no reference to privacy or data subjects at all (e.g., Kubra, FileLocal, and MacroCCS).<sup>116</sup> Some agreements assumed a relationship with the data subject: PayByPhone agreed to “provide an easy to use customer account management website.”<sup>117</sup> Other agreements seemed to assume that the City would remain the point of contact for data subjects: Microsoft, which hosts and processes a variety of City data, committed *not* to respond to data subject requests absent the City’s prior written consent or a legal obligation.<sup>118</sup> There was next to no language obligating vendors to notify data subjects of anything, except in the case of a data breach as discussed in the next section. And long-term retention was, if mentioned, framed as a benefit.

A variety of contracts (e.g., those with CopLogic, Hewitt, and Affirma) addressed the privacy-related concept of “confidential information.” Confidential information does not always intersect with the sensitive information of data subjects.<sup>119</sup> For example, the Motorola agreement defines it as “any information that is . . . marked, designated, or identified at the time of disclosure to [sic] as being confidential.”<sup>120</sup> However, confidential information can so intersect. CopLogic, a software IT company that services the City’s online police reporting system, defines confidential information to include certain “City employee information” such as Social Security numbers or email addresses.<sup>121</sup> Confidential information can also include the vendor’s own “ideas, concepts, know-how or techniques,” i.e.,

---

RECOMMENDATIONS ON DEVELOPING A MEANINGFUL PRIVACY POLICY (May 2014), [https://oag.ca.gov/sites/all/files/agweb/pdfs/cybersecurity/making\\_your\\_privacy\\_practices\\_public.pdf](https://oag.ca.gov/sites/all/files/agweb/pdfs/cybersecurity/making_your_privacy_practices_public.pdf).

116. Kubra, FileLocal, and MacroCCS jointly service the Washington State Business License and Tax Portal Agency, an online portal to pay for business licenses and taxes for several Washington cities including Seattle.

117. PayByPhone Technologies, Inc. Vendor Contract #2992, § 10 “Ownership and Privacy of End User Information,” at 4 (2015) (on file with authors).

118. *See, e.g.*, Microsoft Enterprise Agreement Amendment CTM01E68910, § 9 “Office 365 Security Terms,” Subsection (A) Privacy, at 11 (2013) (on file with authors).

119. For two important discussions of the relationship between privacy and confidentiality, see Neil M. Richards & Daniel J. Solove, *Privacy’s Other Path: Recovering the Law of Confidentiality*, 96 GEO. L.J. 123 (2007), and Woodrow Hartzog, *Reviving Implied Confidentiality*, 89 IND. L.J. 763 (2014).

120. Motorola Solutions, Inc. Blanket Contract 2592, § 32, subsec. 8, at 14–15 (2011) (on file with authors).

121. Coplogic, Inc. Blanket Contract 2708, § 35.2.1, at 21 (2010) (on file with authors).

information proprietary to that business.<sup>122</sup> Where information is designated confidential it may be subject to special protections by agreement, including the prospect of an audit of the vendor to ensure they are processing the information correctly.

Two agreements discussed internal measures to ensure that only the vendor employees who need access to City data would have it—in general, a best practice in consumer privacy. Microsoft committed that “Microsoft personnel will not use, process, or disclose customer data without authorization,” and further that “Microsoft personnel are obligated to maintain the confidentiality of any customer data and this obligation continues even after their engagement ends.”<sup>123</sup> Volgistics, too, provided that “Volgistics customer service employees will have access to customer data as needed for the purpose of answering customer support inquiries,” and also that “Volgistics accounting staff can only see part of your credit card information.”<sup>124</sup> No other contract we sampled limited internal access.

Quite a few agreements mentioned how long information would be retained—a typical subject of privacy policies in the commercial context. Retention terms varied, with longer retention generally framed as a selling point. For example, Socrata, which manages the City’s open data portal, advised it would retain City records for six years after the expiration or termination of the agreement.<sup>125</sup> Socrata also provides that it will keep the data at the same geographic location unless the City authorizes a new location in writing. Other contracts provided for the return of the data. For example, Truven, a health analytics company, committed to “provide to the City all City-owned data, property and deliverable . . . in the format originally sent to the Vendor by the City or its Data Sources.”<sup>126</sup>

Other agreements discussed the conditions under which City data would ever be shared with a third party. For the most part, the relevant language committed the vendor to hold its subcontractors to the same obligations the vendor has to the City. Language such as Oracle’s is common: “Any subcontract made by Vendor shall incorporate by reference

---

122. Affirma Consulting, Agreement Number CRU 2013-002, § 22, subsec. G, at 11 (2013) (on file with authors).

123. *See, e.g.*, Microsoft Enterprise Agreement Amendment CTM01E68910, § 9 “Office 365 Security Terms,” subsec. (A)(e), at 11 (2013) (on file with authors).

124. Volgistics is a company that offers software-based coordination of volunteers, of which the City has many.

125. Socrata, Inc. Blanket Contract 3406, § 27 “Review of Vendor Records,” at 24 (2014) (on file with authors).

126. Truven, Vendor Contract 3150, § 41.7.5 “Termination,” at 22 (2013).

all the terms of this Contract . . . .”<sup>127</sup> Confidential information, however defined, sometimes enjoyed special protection against disclosure.

Several vendor agreements at least contemplated the possibility of sharing with data with third parties. The Acyclica contract reserved the right for the parties to renegotiate data ownership, “specifically with respect to reselling of data,”<sup>128</sup> whereas Truven required the City to *opt out* of sharing its information with Truven’s MarketScan program and, in doing so, give up the “MarketScan contribution discount.”<sup>129</sup> We were unable to determine whether the City decided to participate in MarketScan, and we imagine the data would only be shared in the aggregate in any event.

A noteworthy feature of many of the contracts was the treatment of privacy and security; many contracts did not explicitly address privacy concerns by name even though they did so for security. Privacy and security are both important abstractions governing the use of data but are conceptually distinct enough to warrant separate analysis.

## 2. Security

One of the main concerns of stakeholders—in general, and specifically in our study—is the adequacy of security around data. We are all aware of major breaches affecting even the most sophisticated institutions. Security is one of the venerated Fair Information Practice Principles (FIPPs), which the FTC and others use as a lodestar for privacy policy.<sup>130</sup> A statement of security practices is required for websites operating in California, as alluded to above, and most states impose obligations on data custodians to notify data subjects and the relevant authorities of a breach.<sup>131</sup>

The agreements we sampled and reviewed fared better on security than privacy. Ten out of eighteen specifically reference the adequacy of data security. Several called for security audits or else required vendors to provide documentation of their security policies. Claims of security varied in specificity. For instance, Parkeon simply states it will take “an appropriate

---

127. Oracle America, Inc. Blanket Contract 3025, § 13b, at 4 (2013).

128. Acyclica Attachment to the Western Systems Purchase Order, § 2.6.1, at 2 (on file with authors).

129. Truven, Vendor Contract 3150, exhibit B § 13(g), at 6 (2013) (on file with authors).

130. FED. TRADE COMM’N, PRIVACY ONLINE: FAIR INFORMATION PRACTICES IN THE ELECTRONIC MARKETPLACE (2000), <https://www.ftc.gov/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission>.

131. Forty-seven states have laws on the books governing disclosure of data breaches. For a current list see *Security Breach Notification Laws*, NAT’L CONFERENCE OF STATE LEGISLATURES, <http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx>.

standard of due care,”<sup>132</sup> whereas others offered specific benchmarks. Motorola stated it would treat the city’s data as if it were their own, internal data.<sup>133</sup> PayByPhone pegged its standard to the robust Payment Card Industry Data Security Standard.<sup>134</sup> And CopLogic offered an attestation that a security auditor had tested its system for “common security vulnerabilities.”<sup>135</sup>

Several companies dealt specifically with the important issue of encryption, i.e., storing or communicating information in ways that would ordinarily be unintelligible if accessed or intercepted by an unintended party.<sup>136</sup> Acyclica, a company that collects and processes traffic data, promised that the City’s data would be “encrypted to fully eliminate the possibility of identifying individuals or vehicles.”<sup>137</sup> The health analytics firm Truven specified 128-bit Secure Socket Layer (SSL) encryption of some data.<sup>138</sup> Volgistics also uses SSL for data in transit and storage.<sup>139</sup> Finally, Microsoft uses encryption on data and media that is sent on public networks or leaves its facilities.<sup>140</sup> Acyclica, Truven, and Volgistics also refer to the use of de-identification techniques separate from encryption.

Many states, including Washington, obligate companies that experience data breaches to notify consumers and the authorities within a specified time period.<sup>141</sup> Regardless, parties are free to delineate additional, legally

---

132. Parkeon, Inc. Vendor Contract 1163, Attachment 1 § 5, at 7 (2004) (on file with authors).

133. Motorola Solutions, Inc. Blanket Contract 2592, exhibit A “Data Information Security Services,” at 5 (2011) (on file with authors).

134. PayByPhone Technologies, Inc. Vendor Contract 2992, § 13 Security, “Privacy and Compliance,” at 5 (2015) (on file with authors).

135. CopLogic, Inc. Blanket Contract 2708 § 16 “Security,” at 12 (2010) (on file with authors).

136. We presume many other vendors make routine use of encryption and simply do not mention it.

137. Acyclica Attachment to the Western Systems Purchase Order, § 2.5.1, at 2. This language is probably a little too strong. It may be possible for sophisticated parties to identify people or objects even if encrypted, for instance, by breaking the encryption.

138. Truven, Vendor Contract 3150, exhibit B § 15 “Data Communication,” at 6 (2013) (on file with authors).

139. Volgistics Online Form Security and Privacy Policies, “Security Policies,” at 2 (2015) (on file with authors).

140. EA Amendment CMT01E68910, § 9 “Office 365 Security Terms,” § (D)(a)(v) 4.A., at 14 (2013) (on file with authors). Microsoft encrypts Customer Data that is transmitted over public networks; B. Microsoft restricts access to Customer Data in media leaving its facilities (e.g., through encryption).

141. *See* WASH. REV. CODE § 19.255.010(1). Section 19.255.010(1) states:

Any person or business that conducts business in this state and that owns or licenses computerized data that includes personal information shall

consistent terms in the event of a security breach and often do so. In the documents we analyzed, we noted that a few vendors committed to notifying the City “immediately” (Socrata) or within one business day (Parkeon).<sup>142</sup>

While state laws may obligate companies to disclose breaches, they do not purport to delineate legal responsibility in the event of a breach.<sup>143</sup> We found that specific vendors attempted to contractually absolve themselves of liability should a breach occur. This could occur generally through an arbitration agreement (e.g., Tokusaku) or vendors could absolve liability quite specifically in the event of a breach. For example, Socrata disclaims *all* damages for loss of data, “whether or not resulting from acts of God, communications failure, theft, destruction or unauthorized access to Socrata’s records, programs, or services.”<sup>144</sup> In contrast, still other vendors (e.g., Hewitt and Microsoft), provide for credit monitoring or other “direct damages” in the event of a breach. The City itself could be held accountable consistent with sovereign immunity.<sup>145</sup>

### 3. *Analysis*

The agreements we reviewed were so-called “enterprise” agreements, i.e., made between sophisticated parties. It would not necessarily be fair to judge agreements between cities and firms against consumer privacy policies or terms of use. Thus, we might not expect the agreements to exactly track the Fair Information Practice Principles of notice, access, choice, and security, or to adhere to the strictures of the California Online Privacy Protection Act requiring every website to identify what data it collects and

---

disclose any breach of the security of the system following discovery or notification of the breach in the security of the data to any resident of this state whose unencrypted personal information was, or is reasonably believed to have been, acquired by an unauthorized person.

*Id.*

142. Socrata, Inc. Blanket Contract 3406, subsec. 5.2.8, at 16 (2014) (on file with authors); Parkeon Vendor Contract 1163, Attachment 1, sec. 5 “Security Standards,” at 7 (2004) (on file with authors).

143. *See, e.g.*, WASH. REV. CODE § 19.255.010(1).

144. Socrata, Inc. Blanket Contract 3406, subsec. 17, at 21 (2014) (on file with authors).

145. *See* Kelso v. Tacoma, 390 P.2d 2 (Wash. 1964) (holding that the State of Washington has waived sovereign immunity in tort cases and municipal sovereign immunity); *see also* Locke v. City of Seattle, 172 P.3d 705 (2007). *But see* Cummins v. Lewis County, 133 P.3d 458 (Wash. 2006) (holding that the public duty doctrine still applies to the State of Washington). For a discussion of government liability in Washington see Michael Tardif & Rob McKenna, *Washington State’s 45-Year Experiment in Government Liability*, 29 SEATTLE U. L. REV. 1 (2005).

how it is used and safeguarded,<sup>146</sup> even as we employ these standards as benchmarks of best practice.

More so than an individual consumer, however, the City is in a position to dictate the terms on which it will transact. Many of those terms—such as adequate security—should apply in all of the City’s dealings around resident or City data. What we most clearly observed in the vendor contracts was a lack of standardization. The city reserves very disparate rights against its various vendors, and receives a wide range of positive guarantees. Privacy basics—such as notification requirements, security standards (including encryption), and internal safeguards against unauthorized access—were not specifically delineated in many instances. Companies like Volgistics and Microsoft made extensive mention of privacy and security, laying out exact terms. But other companies made almost no mention of these.

This reflects the status of cities as market makers, not market takers. Law is not the only modality of regulation. Another is markets: cities can and will drive business decisions because they are major potential customers. An insistence that municipal vendors in the data space agree to basic commitments around privacy and security can make city and citizen data more secure all over the country by raising the market bar.

#### IV. RECOMMENDATIONS

The Article thus far has described the expectations around, and inner workings of, Seattle’s open government initiative and other data processes. A final section outlines some tentative recommendations on the basis of what the team has learned. Though researched for the City of Seattle, the practical nature of the seven recommendations shown in this section could be considered valuable to any municipality seeking public trust, privacy, and social justice on the road to open data.

##### A. INVENTORY DATA ASSETS

Our first recommendation involves creating a complete inventory of datasets, the fields within those datasets, and metadata explaining how the information was collected, its purpose and use for the municipality, and any other relevant descriptors concerning the proper management and disposition of the data.

While much of this Article has focused on the contents of datasets, the topic of metadata should not be ignored. Metadata can provide the municipal organizations charged with governing data release with

---

146. CAL. BUS. & PROF. CODE §§ 22575–22579 (West 2014).

information critical to understanding and hopefully acting within the municipal and decidedly public context for the data.<sup>147</sup>

A common standard for metadata is the Dublin Core, a list of categories useful for storing and classifying data. The fifteen fields that comprise the core include: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights.<sup>148</sup> Amongst these categories are many fields for metadata that are potentially valuable for storing, among other things, records that explain the purpose of collecting the data on the part of the responsible department or office, the public uses of the data, a description of the anticipated public benefits of those uses, the classification of the data (e.g., sensitive, critical infrastructure), the nature of the subjects, the sensitivities of the data, restrictions on releases, requirements for aggregation prior to release, suggested qualifications for note in exemption logs in reply to public disclosure requests, a list of the third parties allowed access to the data, the allowable uses or restrictions on use of the data by those third parties, required security measures, applicable regulations, and a note explaining the ex ante and ex post analyses of risk to privacy and social justice conducted in relation to the distribution of the data.

Metadata includes field names. As our technical analysis highlights, municipalities and their related government offices (i.e., counties, special districts, states) should develop and share a data dictionary—a standardized nomenclature for data fields and entries. This tool can provide multiple efficiencies. It can assist departments and the public in interpreting and using municipal data. Departments will find it easier to locate and identify existing information. It can also reduce the chance that work would be unnecessarily duplicated, as would occur if someone found it difficult to find or properly interpret the datasets that already exist.

A more exact and shared naming convention can also reduce the time and effort needed to determine the risk of harm in releasing datasets to the public. In the case of our research, several of our technical strategies were designed to simply deal with the fact that no shared lexicon currently exists for the field names used by municipal departments. “Address” is just as likely to appear as “ADDR,” “Street Address,” and “Location,” and the difference creates unnecessary hurdles for ex ante analysis of risk of release. Any effort

---

147. For a definition of metadata, see KITCHIN, *supra* note 102, at 8.

148. *See About Us*, DUBLIN CORE METADATA INITIATIVE, <http://dublincore.org> (last visited Sept. 1, 2015) (“The Dublin Core Metadata Initiative (DCMI) supports shared innovation in metadata design and best practices across a broad range of purposes and business models.”); *Dublin Core Metadata Element Set*, DUBLIN CORE METADATA INITIATIVE (June 14, 2012), <http://dublincore.org/documents/dces>.

that has to be spent to interpret the existing data is effort that could be saved and spent elsewhere.

B. REQUIRE EACH UNIT TO DEVELOP AND SUBMIT DATA POLICIES

For cities trying to thread the needle of protection for private and social information while enjoying the ability to make other sets of data available to the public, operating as a federated system has its benefits and its drawbacks. Departments in a federated system will have a diversity of strategies that have evolved to implement the policies they have each created and tackle the problems they have each encountered. Revealing the possibility to one department that they may emulate a practice in another may be just the thing to assist departments. In Seattle, for example, some departments appeared to be more comfortable than others sorting meaningful from frivolous examples of public disclosure requests, and denying requests with an explanation filed in their exemption log.

For Seattle, with newly adopted privacy principles, this is an opportune time to learn about the variety of policies departments have already been exercising that, whether they realized it or not, have had the effect of preserving or compromising privacy and social equity. The Department of Information Technology and the Mayor's Office are intent on delivering a citywide privacy policy. The successful implementation of such a policy will depend on the ability of people in these departments to discern the degree to which each department is already delivering practices that preserve privacy and social equity, and to focus attention where it is needed to assist departments that may feel overwhelmed by the shift in priorities.

Consider, in this light, the contrast in notice and consent provided to the residents of Seattle from the Department of Transportation's enlistment of the services of Strava and Acyclica. One need not observe the presence of a field name in a dataset to realize that the data can be used to identify persons. As Montjoye et al. have shown, in their analysis of hourly information flow from devices which record and track the movements of people in time and space by keying in to the MAC address of personal devices (similar to those deployed in Seattle), the traces of mobility left by persons across the urban landscape are highly unique.<sup>149</sup> With only four data points observed in a day, 95% of MAC addresses and persons can be identified. Within the spatial scale of a municipality the task of re-identification is further eased by the classification of municipal land use into residential, office, and other forms of commercial space. Only one, or

---

149. See generally Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, SCI. REP. (Mar. 25, 2013), <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607247/pdf/srep01376.pdf>.

perhaps two data points would be needed to identify most individuals: the location at time of day when statistically likely to be in residence, and the location at time of day when likely to be at school or work.

With these facts in mind, notice and consent would seem to be among the prudent cautionary measures necessary for preserving public trust in the privacy-preserving efforts of the Department. Strava's application does not capture the data flow of the entire population, and as an opt-in program the data has limitations, yet it is data that participants agree to provide and it has proven useful to the Department for the study of travel behavior. Acyclica's data covers more of the population and this fact is due to the lack of notice, choice, and related attention to privacy that accompanied the installation and contractual arrangements for Wi-Fi and Bluetooth sniffers in the public spaces around Seattle. The City can create and test new avenues for notice, consent, and choice. People can opt-out of the program if they are aware of it and capable of following the instructions to do so. The City can also adopt more restrictive policies for permitting the distribution of devices for surveillance in public space.

The next step for the City is to ask how important is the public use for which this data is collected, and who should make this determination? If the public use is deemed valuable enough to the taxpayer (including all ancillary costs envisioned to make the data secure), the next question to ask is how relevant this data is—in its entirety—to the public uses for which it is collected. One can question the need for a sample of this size, the frequency of the collection, the granularity and choice of spatial collection, and of course, retention and distribution of the data. If used, for example, for traffic operations on congested arterial streets, and such use is sanctioned by the public or elected representatives, then the obvious condition that should follow is the limitation of the spatial extent of collection. There is no need for traffic operations to include the monitoring and evaluation of travel behavior in the residential zones of the city, where the ease of personally identifying individuals on the basis of time and location is most likely. Like the black-out dates that airlines have employed to prevent the use of discount travel during peak periods, municipalities should adopt black-out zones, to prevent the use of personally identifying surveillance technologies.

What these two cases suggest is also the extent to which a federated system lends itself to an ad hoc approach to problems that are holistic in nature, such as the problem of analyzing the potential privacy and social equity harms involved in data releases. For this, a governance structure is needed.

### C. ESTABLISH NESTED GOVERNANCE STRUCTURE

Municipalities need structures to more effectively govern the releases of data, via push, pull, and spill. They need governance structures that operate on more than one level, that emulate the need to coordinate and provide some hierarchy to the complex decisions that municipalities must make through the release of data.

A nested governance structure could help municipalities develop citywide policies and avoid ad hoc decision-making. Such a structure could involve oversight from a municipal decision-making body analogous to an Institutional Review Board (IRB), which are convened to review proposed academic research involving human subjects. At the department level, such a structure would include clear guidance about the types of activities that would be exempt from review by the municipal IRB. The activities of interest would span the life cycle of data, to collection, use, retention, deletion, as well as release. Activities that are not exempt would be elevated for review by the IRB.

Emphasizing the importance of informed, meaningful consent, Barocas and Nissenbaum explain that notice and consent are most effectively refined through the services of such a review board.<sup>150</sup> In their explanation, they borrow from the literature on human subject research in medicine, applying these basic insights to the broader case of notice and consent for privacy.<sup>151</sup> They acknowledge that patient interactions take place against a backdrop of trust, and that consent or waiver should be interpreted narrowly. Quoting O'Neill and Manson, they explain that obligations and expectations of medical service providers are not discarded when patients consent. Consent is requested of subjects in limited ways, for limited times and very specific purposes.<sup>152</sup> In consenting to an appendectomy, one does not consent to other incisions, or to incisions by persons other than the relevant surgeon. Furthermore, consent is not required for expected behaviors; it is required for behaviors that depart from what is expected. The burden is on the researcher or clinician to, in giving notice, “describe clearly the violations of norms, standards, and expectations for which a waiver is being asked.”<sup>153</sup> In applying these insights to the more general problem of privacy amidst big data, the authors suggest, “[a] burden is upon the collector and user of data to explain why a subject has good reason to consent, even if consenting to

---

150. See Barocas & Nissenbaum, *supra* note 97, at 64.

151. *Id.* at 44–75.

152. *Id.* at 64–65.

153. *Id.* at 65.

data practices that lie outside the norm. That, or there should be excellent reasons why social and contextual ends are served by these practices.”<sup>154</sup>

In the case of Seattle, we have sought to illustrate the contextual circumstances that surround the municipal rush to big data and open data. Several departments are adopting technologies that collect rich datasets about the people living and working in Seattle. Once collected, the data can be subject to public disclosure request, and may be considered for release to an open data portal. All of these activities can occur in ways that pay scant attention to the potential effect on privacy or social justice from releases of data. For example, the “8 Principles of Open Government Data,” used to structure the review and release of business license data by the Departments of Information Technology and Financial and Administrative Services, were designed for the purposes of promoting the release of data. In this system of reviewing and releasing data, there is no equivalent guidance in practice to safeguard privacy and social justice.

The process of data review and release is devoid of the contextual and subject-oriented privacy protection that Barocas and Nissenbaum define. Practices to safeguard privacy and social justice are, in the current process, reduced to the evaluation of individual fields within isolated datasets. Given this, it is no wonder that public trust in the privacy-preserving actions of municipalities remains suspect. We suggest the adoption of a municipal IRB, tasked with protecting privacy and social justice, with the authority to veto and condition the collection, use, and release of data, and the interdisciplinary capability and experience to evaluate the public interest in such decisions. Given the countervailing interests of open data and privacy, it is worth mentioning that these aims should not be the responsibility of the same person or division within a city department.

IRBs, however, are not needed in every case of review, and the Department of Information Technology may seek to produce a list of datasets and their fields that may be handled through administrative review within the department that owns the data, or exempted from review altogether. Municipal IRBs should be called into service only when the data subjects are employees of the city, residents, or workers. The IRB can be asked to review requests from departments for public release of data to portals or online platforms and any accompanying supportive analysis, such as an analysis of the nexus between the collection of the data, its public uses, the interests of the taxpayer, and privacy and social justice implications. The City should also consider using the IRB to evaluate public disclosure requests that pose privacy or social justice problems, for which there are no

---

154. *Id.* at 67.

clear exemptions in the PRA. This should result in recommendations rendered on a case-by-case basis, yet informed by a body of knowledge of preceding cases and their outcomes, as well as ongoing research in the rapidly moving field of re-identification.

D. ESTABLISH AND DISSEMINATE EX ANTE PROTOCOLS FOR PUSH, PULL, AND SPILL

Cities should plan for the fact that departments may want to release data by pushing it out to public portals when they should not or that departments may inadequately act or invest to prevent the pull or spill of data. One effective way to do this is to establish and disseminate protocols for investigating datasets, in order to educate departments about how to preserve privacy and social equity by curbing or curtailing certain types of releases.

Our suggestions stem from our study of how multiple databases may be joined after they have been published. Possibly the simplest approach a city could take in a protocol to evaluate releases *ex ante* would be to programmatically perform the same kinds of join strategies which our research team did—and perhaps a few others that we did not have time to develop. The join strategies would illustrate the overall joins made possible with other public datasets (and private ones if available) if the proposed new data were to be published. This method would result in two useful artifacts:

1. The resulting joined dataset, which could highlight newly harmful combinations of data made possible with the introduction of new data to the existing corpus of publically available data.
2. A network map that shows precisely which fields would be used to accomplish joins resulting in privacy harm.

The same method could be used to discover and eliminate existing indexical fields, which cause the greatest degree of correlation across the continuum of privacy related attributes in existing datasets. By adopting this practice, and relying on as many existing datasets as possible, the City of Seattle can reduce the likelihood of, and thus manage the risk associated with, the joining of independent datasets in ways which may cause privacy harm.

E. CONDUCT PUBLIC RECORDS AUDIT AND TRAINING

We recommend based on the above that cities engage in audits and training exercises whereby municipalities compare the text of state and federal public records acts with what individual departments are doing on the ground. In the case of Seattle, the City has protocols in place, by

department, on how to respond to PRA requests. However, it is important for *all* employees—not just those with responsibility for responding to outside requests—to understand the law and the City’s interpretation of the law. This will help reduce uncertainty and fear around the prospect of abusive pulls or spills of employee data.

In our engagements with City employees, we noticed variation in the understanding and application of public records requests. First, as noted, not all departments adopted the same posture toward a request for information. Parks and Recreation, which deals mostly with children and families, adopted a relatively restrictive stance.<sup>155</sup> The Police Department had to come up with entirely novel procedures to accommodate massive requests for information in the form of video recordings, and defaulted toward sharing everything (with some modifications for privacy).

We also noticed that employees articulated fears about abusive behaviors that should not have been possible under the text of the PRA. The act provides an exception, for instance, for personal information about an employee.<sup>156</sup> Nevertheless, employees worried that other employees or the public would gain access to information for the purposes of relationships, bias, or embarrassment. When the PRA exception for employee personal information was pointed out in an interview, the room erupted in laughter, as if to suggest the exception would not be honored.<sup>157</sup>

This is not to say that any city should ignore the role of context—it may be a good thing that departments do not all react identically to a request for information. However, there should be some standardization. In particular, all employees involved in responding to public records requests should know the exceptions and the reasons behind them, and generally be able to fall back on a clearly articulated policy.

#### F. EXPLORE CONDITIONED ACCESS OF MUNICIPAL DATA

We recommend that cities explore vehicles by which to make certain data available under specific conditions. This is a fairly common practice. Companies, of course, routinely condition access to information on signing a nondisclosure agreement. In the public sector, more than twenty states condition access to voter databases on noncommercial use.<sup>158</sup> Federal

---

155. Interview, Seattle Parks and Recreation Personnel, Seattle, Wash. (Mar. 5, 2015).

156. WASH. REV. CODE § 42.56.230(3) (2014) (“The following personal information is exempt from public inspection and copying under this chapter: . . . Personal information in files maintained for employees, appointees, or elected officials of any public agency to the extent that disclosure would violate their right to privacy.”).

157. Focus Group, City Employees, in Seattle, WA (Mar. 9, 2015).

158. *See, e.g.*, WASH. REV. CODE § 29A.08.720(2). That section directs:

election law has similar provisions. As cities open up more and more data, they should consider whether one or more use restrictions would be appropriate.

In our focus groups, several citizens and most privacy advocates expressed concern over the prospect that the City would push data for transparency reasons that would instead be used for commercial or political purposes that were disadvantageous to consumers and citizens. Examples included lenders writing off neighborhoods with respect to offers of credit and politicians ignoring complaints from districts with low political participation.<sup>159</sup> There is ample evidence that municipal open data is a major source for data brokers of all kinds.<sup>160</sup> One opportunity might be to follow the example of some states and federal agencies around political data and condition access to certain data sets on noncommercial or nondiscriminatory use. A government might do this when, for instance, citizens may be less likely to participate in a given, beneficial activity such as voting, donating, or volunteering because they fear it will lead them to be targeted for marketing or otherwise cause them to face adverse commercial consequences.

Another example might be conditioning access on the obligation to update the information periodically. The issue here is that commercial entities may copy databases that then become outdated, either because of a mistake (false lien) or because of an update (juvenile record expunged). Meanwhile, although the City now has the correct version, companies and others may be making decisions on the basis of a copy in the hands of a data broker. Presently nothing, apart from industry best practice, obligates these data brokers to keep their databases up to date.

It should be noted that there are a number of pitfalls with this approach. The first is that once data has been released, it is hard to follow. The City

---

The county auditor or secretary of state shall promptly furnish current lists of registered voters in his or her possession, at actual reproduction cost, to any person requesting such information. The lists shall not be used for the purpose of mailing or delivering any advertisement or offer for any property, establishment, organization, product, or service or for the purpose of mailing or delivering any solicitation for money, services, or anything of value. However, the lists and labels may be used for any political purpose.

*Id.* For a summary of state-by-state codes on conditions pertaining to voter list access, see *Voter data use terms and conditions*, NATION BUILDER, <http://nationbuilder.com/voterdata>; see also Kim Zetter, *For Sale: The American Voter*, WIRED (Dec. 11, 2003), <http://archive.wired.com/politics/security/news/2003/12/61543?currentPage=all>.

159. Focus Group, Privacy Activist Organization, in Seattle, Wash. (Feb. 28, 2015).

160. See FED. TRADE COMM'N, *supra* note 9.

might attach rules to its own data but it would have to think through what happens downstream. Imagine, for instance, a condition that commercial users of political data must certify that they will periodically update that data. What if a noncommercial user—a political accountability non-profit—downloads and reposts the data without restrictions? The City would have to look for examples—for instance, in intellectual property licensing—for language that follows the data.

The second is that recent Supreme Court precedent limits the sorts of restrictions that governments can place on uses of data. In *Sorrell v. IMS Health*, the Court invalidated Vermont’s attempt to restrict pharmaceutical companies ability to use doctors’ prescribing history for marketing purposes—a process called “detailing.”<sup>161</sup> The Court found Vermont’s attempt to prevent such targeting to be an unconstitutional restriction on these companies’ speech.

Note that Vermont did not merely condition access to prescription information on using it for a noncommercial purpose. It singled out particular speakers to silence. According to the Court, “Vermont’s law enacts content- and speaker-based restrictions on the sale disclosure, and use of prescriber-identifying information.”<sup>162</sup> Specifically, the Court found that “the statute disfavors specific speakers, namely pharmaceutical manufacturers.”<sup>163</sup> Thus, the Court concluded that the law ran afoul of constitutional prescriptions of discriminating against viewpoints. Had the state instead kept the data itself and released it only on the condition that it not be used for commercial purposes, the Court might not have taken issue.

In general, there may be situations wherein the City wants *some* types of commercial activities—such as the development of a helpful app by a for-profit start up—but would like to avoid others—such as profiling for marketing. These sorts of restrictions are not likely to survive constitutional scrutiny in light of *Sorrell* and other precedent.<sup>164</sup>

#### G. DEVELOP STANDARD VENDOR AGREEMENT

We further recommend that the City of Seattle—and others, as well—create a standard vendor agreement to use as a baseline in all future contracting around City data. This agreement would lay out in clear and simple language the obligations that the vendor takes on by virtue of its custody over City data. These include:

---

161. *Sorrell v. IMS Health, Inc.*, 131 S. Ct. 2653, 2672 (2011).

162. *Id.* at 2663.

163. *Id.*

164. *See id.*; *see also* *Discovery Networks v. City of Cincinnati*, 507 U.S. 410, 424 (1993) (holding that governments may not ban speech merely on the basis that it is commercial).

- maintaining the confidentiality of data subjects;
- restricting access to those within the organization that need it;
- documenting basic digital and physical security;
- specific notification provisions in the event of a security breach;
- specific delineation of responsibility and liability in the event of a security breach; and
- obligations not to share data in any format absent the express consent of the City and/or the data subject, or by required operation of law.

The suggestion is not that the City would use the exact same agreement in each instance. We recognize that department needs will vary on the basis of the task. Moreover, there may be circumstances when the City or a vendor will need to insist on differing terms. Rather, we recommend the development of a baseline reference document such that any departure would have to be specifically justified.

Models for such contracts already exist. For example, Microsoft has a master service agreement around privacy and data security as part of its own vendor toolkit.<sup>165</sup> Moreover, there were specific contracts—in particular, those of Volgistics and Microsoft—that contained much of the recommended language already. And contracts can and do refer to pre-established standards of security such as PCI—which some vendors already mention—and Internal Organization for Standardization and International Electrotechnical Commission 27001 (“ISO 27001”) certification. Ultimately drafting a model agreement may be a task best suited to corporate counsel.

An ancillary, though important, benefit of a standardized vendor agreement would be the effect on the overall market for municipal data. Mid to large-size cities such as Seattle with big information needs and access to considerable resources have the potential to be *market-makers*, i.e., to drive the market toward best practices in privacy and security. Our review of vendor contracts suggests that, with exceptions, the market remains immature in this respect. By insisting on a high bar, the City could not only help justify the trust of stakeholders but improve the overall data ecosystem. We would hope that the City would share any materials it developed with other municipalities.

---

165. See *Supplier Privacy Toolkit*, MICROSOFT, <http://www.microsoft.com/about/companyinformation/procurement/toolkit/en/us/requirements.aspx> (last visited July 21, 2015).

## V. FUTURE WORK

This research was motivated by three central questions: does the City of Seattle's open data initiative increase the public trust in city government; what kind of legal framing could the City use to capture the benefits of open data while addressing legitimate privacy concerns; and what other kinds of harms could arise from government release of data? This Article is a first step, and much work remains to be done.

This case study points toward promising future work in the area of open data research for municipalities and other related governmental entities. Among the research questions raised, we highlight the following:

Municipalities exist to represent and serve the public, and their departments and offices generally share a keen interest in providing benefits to the taxpayer, in the form of efficiencies as well as public goods. If open data does indeed provide taxpayers with an efficient vehicle for transparency and accountability, then there is no reason to question the validity of the movement to open data. And yet, the activities the City recorded in data collection and released for public and perhaps commercial uses were just as likely to focus on residents as they were the government. What public good is served when the names of people given building code violations are made public? What service is improved by publicizing the names of people applying to participate in pea-patch gardening projects? What is the public benefit of tracking the movements of people through their devices across the neighborhoods of the city? In a more striking case, consider police body-worn video. The shocking videos of shootings that raise public attention toward the activities of police capture, often in full view, the officer as well as the suspect. We are shocked in witnessing, during the course of the video, how a suspect becomes a victim. When the body worn video (recorded on cameras that face forward from the chest or shoulder of the officer) provide little more than moving pictures of the residents of the city, one has to ask whether this technology genuinely serves the purposes of transparency and accountability. If the electric eye is observing only one of these parties, what purposes does this fulfill?

If we presume that the rationale for data collection and use is valid, then the question of efficiency comes into focus. As Aaron Wildavsky has noted, efficiency does not tell you where to go, only that you should arrive there with the least effort.<sup>166</sup> On the grounds of efficiency one could question whether the use of advanced information technology—with sensors that detect, discern, and develop thick flows of information in real-time—delivers on its promise of efficiency to the taxpayer. What is the empirical

---

166. AARON B. WILDAVSKY, *SPEAKING TRUTH TO POWER* 131 (1989).

evidence that big municipal data collection followed by big data releases (or big exemptions from releases by State Legislatures), pushed, pulled or spilled, make the City more efficient? When public representatives adopt open data policies, releasing data to the wild, we shift the rules of the game by making private information public. What are the full economic consequences, and how are they distributed amongst the public (who are often the subjects of the data), commercial firms (who often request access to the data about public subjects) and the municipality (whose aim it is to represent the public interest)? What are the distributional consequences—does release heighten or relieve the public of its oft-laden position at the lower end of information asymmetry?

The push, pull, and spill of data from municipalities can predispose the general public and public employees to harms of privacy and social equity. With what legal framework might cities be capable of remedying these harms, and navigating the contested space of data control and release? Much in the case of Seattle may hinge on legal frameworks established by the selective intervention of special interests (public and private) in the adoption of exemptions to the Washington State PRA at the state level, in addition to various privacy-facing federal acts, such as HIPAA. Selective intervention in the rules of the game of state disclosure law suggest that the existing legal framework for balancing privacy and open data is somehow flawed, and this doubt is redoubled through empirically powerful examinations of the inability to use existing legal frameworks—predicated on achieving anonymity by replacing or redacting PII—to protect information that people prefer to keep private. What legal remedies exist, and if they were more widespread, would they be sufficient? What remedies should exist, and how will we know when they are effective?

We've said a lot here; clearly, there is more to be said on the subject.

# TOWARDS A MODERN APPROACH TO PRIVACY-AWARE GOVERNMENT DATA RELEASES

*Micah Altman, Alexandra Wood, David R. O'Brien,  
Salil Vadhan & Urs Gasser<sup>†</sup>*

## ABSTRACT

Governments are under increasing pressure to publicly release collected data in order to promote transparency, accountability, and innovation. Because much of the data they release pertains to individuals, agencies rely on various standards and interventions to protect privacy interests while supporting a range of beneficial uses of the data. However, there are growing concerns among privacy scholars, policymakers, and the public that these approaches are incomplete, inconsistent, and difficult to navigate.

To identify gaps in current practice, this Article reviews data released in response to freedom of information and Privacy Act requests, traditional public and vital records, official statistics, and e-government and open government initiatives. It finds that agencies lack formal guidance for implementing privacy interventions in specific cases. Most agencies address privacy by withholding or redacting records that contain directly or indirectly identifying information based on an ad hoc balancing of interests, and different government actors sometimes treat similar privacy risks vastly differently. These observations demonstrate the need for a more systematic approach to privacy analysis and also suggest a new way forward.

---

DOI: <http://dx.doi.org/10.15779/Z38FG17>

© 2015 Micah Altman, MIT; Alexandra Wood, David R. O'Brien, Salil Vadhan & Urs Gasser, Harvard University.

<sup>†</sup> Micah Altman and Alexandra Wood are the lead authors, with Alexandra Wood creating the initial draft of the manuscript and Micah Altman and Alexandra Wood taking primary responsibility for revisions. All authors, Micah Altman, Urs Gasser, David R. O'Brien, Salil Vadhan, and Alexandra Wood, contributed to the conception of the report (including core ideas and statement of research questions). Micah Altman, David R. O'Brien, and Alexandra Wood were primarily responsible for the methodology (development of the use cases and taxonomies applied), and David R. O'Brien for the project administration. Urs Gasser, David R. O'Brien, and Salil Vadhan contributed to the writing through critical review and commentary. Micah Altman, Urs Gasser, and Salil Vadhan provided scientific direction, and Urs Gasser led funding acquisition. Microsoft Corporation, in collaboration with the Berkeley Center for Law & Technology, supported the research and the writing of this report. In addition, this material is based upon work supported by the National Science Foundation under Grant No. 1237235, the Ford Foundation, and the John D. and Catherine T. MacArthur Foundation. We thank the members of the Privacy Tools for Sharing Research Data project for helpful comments.

In response to these concerns, this Article proposes a framework for a modern privacy analysis informed by recent advances in data privacy from disciplines such as computer science, statistics, and law. Modeled on an information security approach, this framework characterizes and distinguishes between privacy controls, threats, vulnerabilities, and utility. When developing a data release mechanism, policymakers should specify the desired data uses and expected benefits, examine each stage of the data lifecycle to identify privacy threats and vulnerabilities, and select controls for each lifecycle stage that are consistent with the uses, threats, and vulnerabilities at that stage. This Article sketches the contours of this analytical framework, populates selected portions of its contents, and illustrates how it can inform the selection of privacy controls by discussing its application to two real-world examples of government data releases.

### TABLE OF CONTENTS

I.	INTRODUCTION: THE CHANGING LANDSCAPE OF GOVERNMENT RELEASES OF DATA.....	1970
II.	OVERVIEW OF CURRENT PRACTICES FOR RELEASING GOVERNMENT DATA .....	1975
A.	FOUR BROAD CATEGORIES OF GOVERNMENT DATA RELEASES .....	1976
1.	<i>Freedom of Information and Privacy Act Requests</i> .....	1977
a)	Types of Information Released.....	1979
b)	Standards for Making Release Decisions .....	1982
c)	Privacy Interventions in Use .....	1984
2.	<i>Traditional Public and Vital Records</i> .....	1986
a)	Types of Information Released.....	1988
b)	Standards for Making Release Decisions .....	1989
c)	Privacy Interventions in Use.....	1989
3.	<i>Official Statistics</i> .....	1991
a)	Types of Information Released.....	1992
b)	Standards for Making Release Decisions .....	1993
c)	Privacy Interventions in Use.....	1995
4.	<i>E-Government and Open Government Initiatives</i> .....	1997
a)	Types of Information Released.....	1999
b)	Standards for Making Release Decisions .....	2002
c)	Privacy Interventions in Use.....	2004
B.	SHORTCOMINGS IN CURRENT PRACTICES .....	2006
III.	A FRAMEWORK FOR MODERNIZING PRIVACY ANALYSIS .....	2010
A.	CHARACTERIZING PRIVACY CONTROLS, THREATS, VULNERABILITIES, AND USES.....	2011
B.	DEVELOPING A CATALOG OF PRIVACY CONTROLS AND INTERVENTIONS .....	2015
1.	<i>Privacy Controls at the Collection and Acceptance Stage</i> .....	2017
2.	<i>Privacy Controls at the Transformation Stage</i> .....	2020

3.	<i>Privacy Controls at the Retention Stage</i> .....	2023
4.	<i>Privacy Controls at the Release and Access Stage</i> .....	2024
5.	<i>Privacy Controls at the Post-Access Stage</i> .....	2028
C.	IDENTIFYING INFORMATION USES, THREATS, AND VULNERABILITIES.....	2032
1.	<i>Information Uses and Expected Utility</i> .....	2032
2.	<i>Privacy Threats</i> .....	2034
3.	<i>Privacy Vulnerabilities</i> .....	2036
D.	DESIGNING DATA RELEASES BY ALIGNING USE, THREATS, AND VULNERABILITIES WITH CONTROLS .....	2040
1.	<i>Specifying Desired Data Uses and Expected Benefits</i> .....	2041
2.	<i>Selecting Controls</i> .....	2042
IV.	APPLYING THE FRAMEWORK TO REAL-WORLD EXAMPLES OF GOVERNMENT DATA RELEASES .....	2048
A.	PUBLIC RELEASE OF WORKPLACE INJURY RECORDS .....	2049
1.	<i>Collection and Acceptance Stage</i> .....	2049
2.	<i>Retention Stage</i> .....	2051
3.	<i>Post-Retention Transformation</i> .....	2052
4.	<i>Release and Access Stage</i> .....	2052
5.	<i>Post-Access Stage</i> .....	2056
6.	<i>Aligning Uses, Threats, and Vulnerabilities with Controls</i> .....	2056
B.	MUNICIPAL OPEN DATA PORTALS.....	2059
1.	<i>Collection and Acceptance Stage</i> .....	2060
2.	<i>Retention Stage</i> .....	2061
3.	<i>Post-Retention Transformation</i> .....	2061
4.	<i>Release and Access Stage</i> .....	2063
5.	<i>Post-Access Stage</i> .....	2067
6.	<i>Aligning Use, Threats, and Vulnerabilities with Controls</i> .....	2068
V.	SUMMARY.....	2070

## I. INTRODUCTION: THE CHANGING LANDSCAPE OF GOVERNMENT RELEASES OF DATA

Transparency is a fundamental principle of democratic governance. Making government data more widely available promises to enhance organizational transparency, improve government functions, encourage civic engagement, support the evaluation of government decisions, and ensure accountability for public institutions. Releases of government data also promote growth in the private sector by guiding investment and other commercial decisions, supporting innovation in the technology sectors, and promoting economic development and competition broadly.<sup>1</sup> Furthermore, improving access to government data also advances the state of research and scientific knowledge, changing how researchers approach their fields of study and enabling them to ask new questions and gain better insights into human behaviors.<sup>2</sup> For instance, the increased availability of large-scale datasets is advancing developments in computational social science, a field that is rapidly changing the study of humans, human behavior, and human institutions, and effectively shifting the evidence base of social science.<sup>3</sup> Scientists are also developing methods to mine and model new data sources and big data, and data collected from people and institutions have proven useful in unexpected ways. In the area of public health, Google Flu Trends, which provides a useful and timely supplement to conventional flu tracking methods by analyzing routine Google search queries, is a widely publicized example of the unexpected

---

1. See generally REGINA POWERS & DAVID BEEDE, U.S. DEPT' OF COMMERCE, FOSTERING INNOVATION, CREATING JOBS, DRIVING BETTER DECISIONS: THE VALUE OF GOVERNMENT DATA (2014) (discussing the many benefits of government releases of data).

2. See Micah Altman & Kenneth Rogerson, *Open Research Questions on Information and Technology in Global and Domestic Politics—Beyond “E-,”* 41 PS: POL. SCI. & POL. 835 (2008); Gary King, *Ensuring the Data-Rich Future of the Social Sciences*, 331 SCIENCE 719 (2009); David Lazer et al., *Computational Social Science*, 323 SCIENCE 721 (2009).

3. See sources cited *supra* note 2.

secondary uses of data.<sup>4</sup> These are, of course, just a few examples of the many benefits of opening up access to data.<sup>5</sup>

For these and related reasons, governments and civic advocates are increasingly recommending that open access be the “default state” for information collected by government agencies.<sup>6</sup> This rationale drives the open government initiatives launched in recent years by federal, state, and municipal governments to release large quantities of information, much of which is about individuals, to the public through a variety of channels.<sup>7</sup> These programs encourage agencies to adopt a presumption of openness, to the extent the law allows, and publish information online in open formats that can be accessed and processed through a variety of applications.<sup>8</sup>

However, a major challenge for any public release of data about individuals is providing meaningful protection of privacy interests.<sup>9</sup> While

---

4. See, e.g., Samantha Cook et al., *Assessing Google Flu Trends Performance in the United States During the 2009 Influenza Virus A (H1N1) Pandemic*, PLOS ONE, Aug. 2011, <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023610>; Justin R. Ortiz et al., *Monitoring Influenza Activity in the United States: A Comparison of Traditional Surveillance Systems with Google Flu Trends*, PLOS ONE, Apr. 2011, <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018687>; N. Wilson et al., *Interpreting “Google Flu Trends” Data for Pandemic H1N1 Influenza: The New Zealand Experience*, EUROSURVEILLANCE, Nov. 5, 2009, <http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19386>.

5. A number of scholars are currently writing about the benefits of open data systems. See, e.g., JOEL GURIN, *OPEN DATA NOW: THE SECRET TO HOT STARTUPS, SMART INVESTING, SAVVY MARKETING, AND FAST INNOVATION* (2014).

6. See, e.g., Exec. Order No. 13,642, 3 C.F.R. 244 (2014) (Making Open and Machine Readable the New Default for Government Information), <https://www.gpo.gov/fdsys/pkg/CFR-2014-title3-vol1/pdf/CFR-2014-title3-vol1-eo13642.pdf>.

7. Paul M. Schwartz, *Privacy and Participation: Personal Information and Public Sector Regulation in the United States*, 80 IOWA L. REV. 553 (1995); Harlan Yu & David G. Robinson, *The New Ambiguity of “Open Government,”* 59 UCLA L. REV. DISCOURSE 178 (2012).

8. E.g., PETER R. ORSZAG, OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, M-10-06, MEMORANDUM ON OPEN GOVERNMENT DIRECTIVE (Dec. 8, 2009), [http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda\\_2010/m10-06.pdf](http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf).

9. Throughout this Article, we use “privacy” and “confidentiality” as generally inclusive and approximately synonymous terms. Note, however, that these terms may have narrower definitions within fields, and such definitions are inconsistent and sometimes conflicting. For example, the statistical disclosure limitation literature defines “privacy” as the right of data subjects to control the manner and extent of sharing of their information and “confidentiality” as the duty of data holders to prevent unauthorized disclosure after collection. See, e.g., Stephen E. Fienberg, *Confidentiality and Disclosure Limitation*, 1 ENCYCLOPEDIA SOC. MEASUREMENT 463 (2005). In contrast, the literature on cryptography often uses “privacy” to refer to controls over disclosure or to

governments are generally required to consider the legal and ethical implications of publicly releasing information about individuals, the lack of an effective legal and regulatory framework for privacy arguably hinders the disclosure and reuse of privacy-sensitive data in practice.<sup>10</sup> Privacy laws and policies can be circumstantial, open to interpretation, and ill-suited to apply at scale.<sup>11</sup> Laws and policies concerning the disclosure of government information are context specific, varying substantially based on the type of information released, the agency releasing it, and the mechanism of release.<sup>12</sup> Most states lack “omnibus data protection laws” and have “scattered laws [that] provide only limited protections for personal information in the public sector.”<sup>13</sup>

Executive agencies frequently release government information under sunshine laws such as the Freedom of Information Act (FOIA),<sup>14</sup> which requires disclosures in response to public records requests provided that no law prohibits the release. Individual agencies retain discretionary authority to withhold or redact certain records that implicate one of a limited set of concerns such as privacy, with most agencies releasing records that have been redacted of directly identifying pieces of information such as names, addresses, dates of birth, and Social Security numbers. In contrast, federal statistical agencies must comply with complex laws and policies that regulate the format of the information to be released, require practices that enhance data integrity and accuracy, and mandate strict confidentiality protections.<sup>15</sup> These agencies use statistical disclosure limitation

---

the absence of a privacy breach. *See, e.g.*, Cynthia Dwork, *Differential Privacy*, in *ENCYCLOPEDIA OF CRYPTOGRAPHY AND SECURITY* 338 (Henk C.A. von Tilborg & Sushil Jajodia, 2d ed. 2011). Furthermore, the information security literature uses the term “confidentiality” to refer to controls over disclosure but in the narrower context of an information system. *See, e.g.*, RICK LEHTINEN, DEBORAH RUSSELL & G.T. GANGEMI SR., *COMPUTER SECURITY BASICS* 197 (2d ed. 2006).

10. *See generally* Paul Schwartz & Daniel Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814 (2011) (describing the inadequacy of a U.S. legal framework that largely rests on a flawed concept of “personally identifiable information”).

11. *Id.*; Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321 (1992).

12. *See* discussion *infra* Part II. Note that, while this Article focuses on government releases of data within the United States, legal frameworks in other countries also lead to inconsistent data release practices across government agencies. For a discussion of these issues in both the United States and Europe, see GEORG AICHHOLZER & HERBERT BURKERT, *PUBLIC SECTOR INFORMATION IN THE DIGITAL AGE* (2004).

13. Schwartz, *supra* note 7, at 605.

14. Freedom of Information Act, 5 U.S.C. § 552 (2012).

15. Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), 44 U.S.C. § 3501 note (2012).

techniques to aggregate information from many individuals, suppress individual-level details, or perturb individual data points in ways intended to mitigate privacy concerns while supporting accurate analyses.<sup>16</sup>

As numerous commentators have shown, ineffective treatment of information privacy and security has become a major stumbling block to efficient access to and use of data.<sup>17</sup> Assessment of privacy risk should encompass the range of threats to privacy; the vulnerabilities that exacerbate those threats; the likelihood of disclosure of information given those threats and vulnerabilities; and the extent, severity, and likelihood of harms arising from those disclosures.<sup>18</sup> Yet privacy risks and harms are difficult to predict because data are accumulated, combined, and used in a wide variety of contexts,<sup>19</sup> and data release programs often fail to address the full range of risks identified within the scientific literature on privacy. There are many examples of individuals being identified in datasets despite the data having been de-identified using traditional techniques such as removing or generalizing sensitive fields.<sup>20</sup> In addition, such techniques significantly reduce the utility of data.<sup>21</sup> On the whole, robust

---

16. FED. COMM. ON STATISTICAL METHODOLOGY, STATISTICAL POLICY WORKING PAPER 22 (SECOND VERSION), REPORT ON STATISTICAL DISCLOSURE LIMITATION METHODOLOGY (2005), <https://fcsml.sites.usa.gov/files/2014/04/spwp22.pdf>.

17. *See, e.g.*, NAT'L RESEARCH COUNCIL, EXPANDING ACCESS TO RESEARCH DATA: RECONCILING RISKS AND OPPORTUNITIES (2005); NAT'L RESEARCH COUNCIL, PUTTING PEOPLE ON THE MAP: PROTECTING CONFIDENTIALITY WITH LINKED SOCIAL-SPATIAL DATA (2007) [hereinafter NAT'L RESEARCH COUNCIL, PUTTING PEOPLE ON THE MAP]; NAT'L RESEARCH COUNCIL, BEYOND THE HIPAA PRIVACY RULE: ENHANCING PRIVACY, IMPROVING HEALTH THROUGH RESEARCH (2009); NAT'L RESEARCH COUNCIL, CONDUCTING BIOSOCIAL SURVEYS: COLLECTING, STORING, ACCESSING, AND PROTECTING BIOSPECIMENS AND BIODATA (2010).

18. *See* discussion *infra* Part III.

19. *See* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010); Ira Bloom, *Freedom of Information Laws in the Digital Age: The Death Knell of Information Privacy*, 12 RICH. J.L. & TECH., Article No. 9 (2006); Amanda Conley, Anupam Datta, Helen Nissenbaum & Divya Sharma, *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV. 772 (2012); Teresa Scassa, *Privacy and Open Government*, 6 FUTURE INTERNET 397 (2014), <http://www.mdpi.com/1999-5903/6/2/397/pdf>.

20. *See, e.g.*, Latanya Sweeney, *Matching Known Patients to Health Records in Washington State Data* (Data Privacy Lab, White Paper No. 1089-1, 2013), <http://www.dataprivacylab.org/projects/wa/1089-1.pdf>; Amitai Ziv, *Israel's "Anonymous" Statistics Surveys Aren't So Anonymous*, HAARETZ (Jan. 7, 2013), <http://www.haaretz.com/news/national/israel-s-anonymous-statistics-surveys-aren-t-so-anonymous-1.492256>.

21. *See, e.g.*, Jon P. Daries et al., *Privacy, Anonymity, and Big Data in the Social Sciences*, QUEUE, Aug. 14, 2014, <https://queue.acm.org/detail.cfm?id=2661641>.

de-identification of individual-level data by traditional statistical disclosure limitation techniques is quite difficult, often provides limited or no real-world privacy protection, and narrows the scope of possible uses of the data.<sup>22</sup> These issues are widely recognized and at the center of current academic and policy discussions about how to balance the privacy risks and utility of de-identified data when sharing it with third parties.<sup>23</sup>

These and related challenges indicate that a more sophisticated approach to data releases is needed to provide strong privacy protection for individuals and to improve the utility of data made publically available.<sup>24</sup> By aggregating data, emerging privacy-aware techniques such as synthetic data, data visualizations, interactive mechanisms, and multiparty computations can offer both better privacy and utility in certain contexts.<sup>25</sup> Yet current laws and policies do not provide much guidance to agencies regarding the implementation of newly emerging privacy protections in their public releases of data.<sup>26</sup> Taken together, the laws, policies, and practices compelling and constraining government releases of information often create uncertainty, discourage data sharing, and fail to adequately protect privacy.

This Article provides an overview of current practices for releasing government data and identifies gaps and inconsistencies in the handling of personal information. To begin to address these issues, it outlines a framework for a modern privacy analysis that takes advantage of recent

---

22. See, e.g., Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 U.C.L.A. L. REV. 1701 (2010).

23. See, e.g., Ann Cavoukian & Khaled El Emam, Info. & Privacy Comm'r of Ontario, Canada, *Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy*, in PRIVACY BY DESIGN 227 (2011); Ohm, *supra* note 22; Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117 (2013); Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011).

24. See Letter from Salil Vadhan, Vicky Joseph Professor of Computer & Applied Mathematics, Harvard Univ., et al. to Dep't of Health & Human Servs. et al., Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections (Oct. 26, 2011), <http://privacytools.seas.harvard.edu/files/commonruleanprm.pdf>.

25. See *id.*; see also Satkartar K. Kinney et al., *Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database* (Ctr. for Econ. Studies Discussion Paper CES-WP-11-04, 2011), <http://www.census.gov/ces/pdf/CES-WP-11-04.pdf>; Ashwin Machanavajjhala et al., *Privacy: Theory Meets Practice on the Map*, 24 IEEE INT'L CONF. ON DATA ENGINEERING 277 (2008), <http://www.cse.psu.edu/~dkifer/papers/PrivacyOnTheMap.pdf>.

26. See, e.g., U.S. GENERAL ACCOUNTING OFFICE, GAO-01-126SP, RECORD LINKAGE AND PRIVACY: ISSUES IN CREATING NEW FEDERAL RESEARCH AND STATISTICAL INFORMATION 105 (2001), <http://www.gao.gov/new.items/d01126sp.pdf>.

advances in data privacy from disciplines including computer science,<sup>27</sup> statistics,<sup>28</sup> and law,<sup>29</sup> and considers the nuances of dealing with different types of data and finely matching privacy controls to the intended uses, threats, and vulnerabilities of a release. This framework provides broad guidance for a systematic analysis. Although the state of the art provides no silver bullets and precludes a mechanistic approach to privacy, it does offer many promising new interventions. We catalog these proposed interventions and offer a framework for selecting feasible ones across all stages of the information lifecycle, from collection through post-access, for the design of a privacy-aware data release.

## II. OVERVIEW OF CURRENT PRACTICES FOR RELEASING GOVERNMENT DATA

Federal and state governments release information to the public through a wide variety of mechanisms that reflect the distinct actors, objectives, legal and regulatory contexts, and institutional capacities at play in each setting. Some releases are made pursuant to requests for records. For instance, federal, state, and municipal government agencies frequently release information in response to freedom of information requests, made under FOIA<sup>30</sup> or a corresponding state law. Governments also release information through registries, available to the public online or in-person at a local government office, which serve important functions such as providing evidence of births, deaths, marital status, and property ownership. Through official statistical records, such as those produced by the Census Bureau and the Bureau of Labor Statistics, governments analyze and disseminate essential statistics related to the American population and economy. In recent years, e-government and open data laws and policies have emerged as the latest mechanisms of release. Federal, state, and municipal governments are implementing such programs and triggering the rapid release of large quantities of data for online inspection or download by the public.

---

27. See, e.g., Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMM. ACM 86 (2011); Ori Heffetz & Katrina Ligett, *Privacy and Data-Based Research*, 28 J. ECON. PERSP. 75 (2014); Erica Klarreich, *Privacy by the Numbers: A New Approach to Safeguarding Data*, QUANTA MAG. (Dec. 10, 2012), <https://www.quantamagazine.org/20121210-privacy-by-the-numbers-a-new-approach-to-safeguarding-data>.

28. See, e.g., Machanavajjhala et al., *supra* note 25.

29. See, e.g., Wu, *supra* note 23.

30. See 5 U.S.C. § 552.

Government agencies attempt to protect the privacy of individuals whose information may be present in these data releases. For example, an agency might redact certain identifiers such as first and last names or might withhold the release of a record entirely due to privacy concerns. In some cases an agency is bound by regulations requiring strong confidentiality protections for collecting and releasing information about individual respondents,<sup>31</sup> while in other cases an agency may be granted wide discretion in making release decisions. Regulatory requirements and the choice of release mechanism often dictate the agency's approach to privacy. However, in light of trends towards openness of data, governments are facing challenges that call for a more nuanced and systematic approach to releasing data.

#### A. FOUR BROAD CATEGORIES OF GOVERNMENT DATA RELEASES

To provide an overview of the range of current practice, we conducted a broad literature review of academic articles and government publications describing releases of information about individuals by federal, state, and local agencies, and the laws and policies governing such releases. An iterative analysis of the releases suggested classifying them into four broad categories<sup>32</sup>: responses to freedom of information and Privacy Act<sup>33</sup> requests,<sup>34</sup> traditional public and vital records,<sup>35</sup> official government statistics,<sup>36</sup> and e-government and open government initiatives.<sup>37</sup> These categories are not meant to be exclusive. For example, a release of data in an open data initiative typically relies on a freedom of information law as the legal justification for the release, so aspects of both the freedom of information and the open government categories will apply to the analysis of such a release. This Article uses the broad categories introduced in this Part, as well as specific cases of data releases within these categories, to explore approaches adopted by governments, associated challenges and shortcomings, and potential ways in which current practices might be improved.

---

31. *See, e.g.*, Confidential Information Protection and Statistical Efficiency Act of 2002, 44 U.S.C. § 3501 note (2012).

32. This categorization excludes data that does not describe humans or human activities. It also excludes information that is not directly collected or managed by government, even if it concerns government actors. For example, the privacy of tweets of government officials is outside the scope of this classification scheme.

33. Privacy Act of 1974, 5 U.S.C. § 552a (2012).

34. *See infra* Section II.A.1.

35. *See infra* Section II.A.2.

36. *See infra* Section II.A.3.

37. *See infra* Section II.A.4.

### 1. *Freedom of Information and Privacy Act Requests*

Governments are required by law to routinely make certain information available to the public. One way they do this is by responding to requests for information submitted pursuant to the Freedom of Information Act, the Privacy Act, and various complementary federal and state laws commonly known as freedom of information or “sunshine” laws.<sup>38</sup> In combination, these laws are intended to strike a balance between the public’s right to know what information the government holds and the government’s interest in safeguarding sensitive information, the release of which could harm protected individual, commercial, or governmental interests.<sup>39</sup>

The Freedom of Information Act was enacted in 1966 to promote transparency and accountability in government, enabling the public to review information collected using public funds and examine the data upon which many policymaking decisions are made.<sup>40</sup> The FOIA process is used very frequently, with requests across all federal agencies totaling 714,231 in 2014.<sup>41</sup> FOIA empowers any person, including a non-citizen, to obtain copies of records held by federal executive agencies by following a simple request procedure.<sup>42</sup> FOIA does not require the requester to specify a purpose or public interest justification; indeed, a majority of FOIA requests are made by businesses for commercial reasons.<sup>43</sup> Similarly, state freedom of information laws generally do not permit agencies to restrict access to information based on the purpose of a request, and various state courts have held that doing so would be impermissible unless

---

38. See, e.g., Government in the Sunshine Act, 5 U.S.C. § 552b (2012) (requiring agency meetings to be open to the public unless covered by a specific exception); Classified National Security Information, Exec. Order 12,958, 3 C.F.R. 333 (1996) (prescribing rules for “classifying, safeguarding, and declassifying national security information”).

39. See John Badger Smith, Comment, *Public Access to Information Privately Submitted to Government Agencies: Balancing the Needs of Regulated Businesses and the Public*, 57 WASH. L. REV. 331 (1982).

40. See Fred H. Cate et al., *The Right to Privacy and the Public’s Right to Know: The “Central Purpose” of the Freedom of Information Act*, 46 ADMIN. L. REV. 41 (1994).

41. U.S. Dept. of Justice, *FOIA Data at a Glance—FY 2009 Through FY 2014*, FOIA.GOV, <http://www.foia.gov/index.html> (last visited Apr. 23, 2015).

42. See, e.g., U.S. Dep’t of Justice v. Reports Comm. for Freedom of the Press, 489 U.S. 749 (1989).

43. See Cate et al., *supra* note 40, at 65; Patricia M. Wald, *The Freedom of Information Act: A Short Case Study in the Perils and Paybacks of Legislating Democratic Values*, 33 EMORY L.J. 649, 665–66 (1984).

authorized by statute.<sup>44</sup> By default, all responsive records must be disclosed upon request unless an applicable exemption, such as privacy,<sup>45</sup> applies. FOIA does not require agencies to notify any person whose information is to be released, nor does it give such an individual an opportunity to contest the disclosure. At the state level, there are limited circumstances under which individuals are entitled to shield their personal information from public release in response to a freedom of information request. One such example is a New York state law that grants handgun permit holders the right to opt out of the disclosure of their personal information under the freedom of information law if they submit an application and an attestation of concerns about personal safety or harassment related to the release of such information.<sup>46</sup> Some state freedom of information laws also expressly allow victims of crimes to shield their personal information from release.<sup>47</sup> Apart from these narrow exceptions, the burden of protecting an individual's privacy interests generally rests with the agency holding the information, rather than with the individual subject of the data. Once released, the information can be used for any purpose and freely disseminated, and no efforts are made to monitor access to the data or mitigate threats to privacy post-release. FOIA specifies penalties for government employees who fail to release information that is required to be released, but there are no penalties for releasing information that should not have been released.<sup>48</sup>

A companion law, the Privacy Act of 1974,<sup>49</sup> may compel or bar disclosure of records sought under FOIA. The Privacy Act generally prohibits federal executive agencies from disclosing personal information about U.S. citizens and legal permanent residents maintained in a system

---

44. *See, e.g.*, *Dunhill v. Director, D.C. Dep't of Transp.*, 416 A.2d 244 (D.C. 1980) (holding that the department of motor vehicles could not deny a marketer of personal information access to the contact information of drivers permit holders because such a denial was not authorized by the statute); *In re Crawford*, 194 F.3d 954 (9th Cir. 1999) (holding that unrestricted access to bankruptcy information, including Social Security numbers, in judicial records "fosters confidence among creditors regarding the fairness of the bankruptcy system" and therefore should be ensured despite the heightened risk of fraud and identity theft).

45. 5 U.S.C. § 552(b)(6)–(7).

46. N.Y. PEN. L. § 400.00(5)(b) (2014); *see, e.g.*, *Erie County Clerk, NYS Firearms License Request for Public Records Exemption* (Apr. 28, 2015), [http://www2.erie.gov/clerk/sites/www2.erie.gov.clerk/files/uploads/FOIL\\_Exemption\\_Form.pdf](http://www2.erie.gov/clerk/sites/www2.erie.gov.clerk/files/uploads/FOIL_Exemption_Form.pdf).

47. *See, e.g.*, CAL. GOV'T CODE § 6254(f)(2) (West 2015).

48. 5 U.S.C. § 552(a)(4).

49. § 552a.

of records, except as authorized by the data subject.<sup>50</sup> It authorizes FOIA-mandated disclosures,<sup>51</sup> but if a FOIA exemption applies, an agency must cite a corresponding Privacy Act exemption and either withhold the records or release them with discretion.<sup>52</sup> The Privacy Act also enables a data subject to access, review, and correct her information in government databases, unless an exemption applies.<sup>53</sup> An individual may submit a written Privacy Act request to access records about herself.<sup>54</sup> An agency cannot deny a first party request unless exemptions to both the Privacy Act and FOIA apply. If an agency maintains an inaccurate record, fails to correct a record upon request, or otherwise fails to comply with the Privacy Act in a way that adversely affects an individual, she may bring a civil action against the agency.<sup>55</sup>

#### a) Types of Information Released

Freedom of information requests, appeals, and litigation have prompted the release of raw data from administrative and oversight records, studies by government agencies, and studies supported by public grants. For example, FOIA litigation led to the 2009 release of data from a National Highway Traffic Safety Administration study on the safety risks of operating a cellphone while driving, and consumer groups subsequently published the data online for public review.<sup>56</sup> The Centers for Medicare and Medicaid Services disclosed payments made by

---

50. § 552a(b). A system of records is defined as “a group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual,” § 552a(a)(5), and the term “record” is defined as “any item, collection, or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a photograph,” § 552a(a)(4).

51. § 552a(b)(2).

52. *See Savada v. U.S. Dep’t of Def.*, 755 F. Supp. 6, 9 (D.D.C. 1991) (“If an individual is entitled to a document under FOIA and the Privacy Act, to withhold this document an agency must prove that the document is exempt from release under *both* statutes.”) (emphasis in original) (citing *Martin v. Office of Special Counsel, Merit Sys. Prot. Bd.*, 819 F.2d 1181, 1184 (D.C. Cir. 1987) (“If a FOIA exemption covers the documents, but a Privacy Act exemption does not, the documents must be released under the Privacy Act . . . .”)).

53. § 552a(d)(1)–(2).

54. § 552a(d)(1).

55. § 552a(g)(1).

56. *See* Matt Richtel, *U.S. Withheld Data on Risks of Distracted Driving*, N.Y. TIMES (July 21, 2009), <http://www.nytimes.com/2009/07/21/technology/21distracted.html>.

pharmaceutical companies to individual doctors and the brand names and quantities of medications the doctors prescribed, and a watchdog group published the data in online searchable databases, along with visualizations and investigative commentary on prescribing patterns and signs of fraud.<sup>57</sup> During a public debate about gun control legislation, a newspaper used freedom of information requests to obtain county agencies' data, from which the newspaper created a widely publicized interactive map showing the names and addresses of handgun permit holders.<sup>58</sup> FOIA also serves as a disclosure mechanism for other laws mandating release of government information. For instance, researchers engaged in federally funded research are required to share data with the sponsoring agency so that the agency can disseminate data produced by the research in response to FOIA requests.<sup>59</sup>

FOIA exempts the following records, among several other categories, from mandatory release: classified records; internal personnel records and agency memos; confidential trade secret or financial information; medical or other similar files, broadly interpreted,<sup>60</sup> that would constitute an unwarranted invasion of privacy; and law enforcement records.<sup>61</sup> Agencies are permitted but not required to withhold or redact records that fall within one of the exemptions,<sup>62</sup> and they are generally encouraged to release exempted information, when possible, "as a matter of good public

---

57. See Lena Groeger et al., *Dollars for Docs: How Industry Dollars Reach Your Doctors*, PROPUBLICA, <https://projects.propublica.org/docdollars> (last updated July 1, 2015) (database of payments to doctors); Jeff Larson et al., *Prescriber Checkup: The Doctors and Drugs in Medicare Part D*, PROPUBLICA, <http://projects.propublica.org/checkup> (last updated June 10, 2015) (database of prescriptions).

58. Gun owners vigorously objected to the publication of this map, and the newspaper replaced the interactive map showing specific addresses with a static high-level map less than a month later when the state legislature passed a law allowing permit holders to request that their personal information be shielded from release under the state freedom of information law. See Dwight R. Worley, *The Gun Owner Next Door: What You Don't Know about the Weapons in Your Neighborhood*, THE JOURNAL NEWS (White Plains, N.Y.) (Dec. 23, 2012), <http://www.lohud.com/apps/pbcs.dll/article?AID=2012312230056>; *LoHud Removes Controversial Gun Owners Map*, NBC N.Y. (Jan. 18, 2013), <http://www.nbcnewyork.com/news/local/Journal-News-Removes-Pistol-Permit-Database-Gun-Owners-Rockland-Westchester-187525461.html> (including a video showing the features of the original interactive map).

59. Shelby Amendment, Pub. L. No. 105-277, div. A, tit. III, 112 Stat. 2681, 2681-495 (1999).

60. U.S. Dep't of State v. Washington Post Co., 456 U.S. 595, 599-603 (1982).

61. 5 U.S.C. § 552b(c).

62. See *Chrysler Corp. v. Brown*, 441 U.S. 281, 293-94 (1979) (holding that the legislative history "support[s] the interpretation that the [FOIA] exemptions were only meant to permit the agency to withhold certain information, and were not meant to mandate nondisclosure").

policy.”<sup>63</sup> State freedom of information laws also sometimes contain an explicit presumption in favor of disclosure. For instance, the California Public Records Act permits an agency to withhold a record only as expressly exempted by the Act or if “on the facts of the particular case the public interest served by not disclosing the record clearly outweighs the public interest served by disclosure of the record.”<sup>64</sup>

The Privacy Act prohibits the release of records, maintained by a federal agency in a system of records, containing any “information about an individual that includes an individual identifier,” which refers to “any element of data (name, number) or other descriptor (finger print, voice print, photographs) which can be used to identify an individual” and includes “as little as one descriptive item about an individual.”<sup>65</sup> Federal courts have applied different tests for determining whether a particular piece of information falls within this definition,<sup>66</sup> and many government records about individuals are not covered. Where it applies, an agency must have written consent to release the information; implied or open-ended consent is insufficient.<sup>67</sup> However, an agency may release such records without consent under twelve enumerated exemptions, which enable disclosures to the Census Bureau, law enforcement agencies, Congress, and consumer reporting agencies, among other recipients.<sup>68</sup> An agency may also disclose information for any “routine use” that is “compatible” with its purpose for collecting the information.<sup>69</sup>

---

63. U.S. Attorney General, Memorandum for Heads of Departments and Agencies Re: The Freedom of Information Act (Oct. 4, 1993), <http://www.justice.gov/oip/blog/foia-update-attorney-general-renos-foia-memorandum>.

64. CAL. GOV'T CODE § 6255 (West 2015).

65. Responsibilities for the Maintenance of Records About Individuals by Federal Agencies, 40 Fed. Reg. 28,948, 28,951–52 (July 9, 1975).

66. Compare *Quinn v. Stone*, 978 F.2d 126, 133 (3d Cir. 1992) (interpreting the term “record” to “encompass[] any information about an individual that is linked to that individual through an identifying particular” and to not be “limited to information which taken alone directly reflects a characteristic or quality” (emphasis omitted)), with *Boyd v. U.S. Sec'y of the Navy*, 709 F.2d 684, 686 (11th Cir. 1983) (holding that information “must reflect some quality or characteristic of the individual involved” in order to qualify as a “record”).

67. “At a minimum, the consent clause should state the general purposes for, or types of recipients [to,] which disclosure may be made.” Responsibilities for the Maintenance of Records About Individuals by Federal Agencies, 40 Fed. Reg. at 28,954.

68. 5 U.S.C. §§ 552a(b)(4), (b)(7), (b)(9), (b)(12).

69. § 552a(b)(3).

Commentators have argued that this provision effectively enables disclosure with very little restriction.<sup>70</sup>

b) Standards for Making Release Decisions

In determining whether information is exempted from mandatory disclosure under freedom of information laws, agencies balance the public interest of disclosure against individuals' privacy interests. Standards guiding this balancing have developed through judicial opinions. The Supreme Court has held that the public interest in disclosure outweighs privacy interests except where disclosures "constitute 'clearly unwarranted' invasions of personal privacy"<sup>71</sup> and where the threats to privacy are "more palpable than mere possibilities."<sup>72</sup> Yet it has also held that records need not contain "highly personal" information or "intimate details" to be considered to be privacy-sensitive.<sup>73</sup> In addition, it has established a "central purpose" test that directs agencies to release information about official government activities but not personally identifiable information that is "intended for or restricted to the use of a particular person or group or class of persons, not freely available to the public."<sup>74</sup> For example, it did not require the State Department to disclose the names of Haitian nationals who had been interviewed by the U.S. government, since such disclosure could subject them to "retaliatory action" and "embarrassment in their social and community relationships."<sup>75</sup> It has found that non-union employees have "some nontrivial privacy interest in nondisclosure" and "in avoiding the influx of union-related mail, and, perhaps, union-related telephone calls or visits, that would follow disclosure" of their home addresses to a trade union.<sup>76</sup>

State agencies sometimes interpret the privacy exemption standard to weigh strongly in favor of withholding or redacting records and improperly refuse to release personally identifiable information.<sup>77</sup> In one

---

70. See, e.g., Robert Gellman, *Does Privacy Law Work?*, in *TECHNOLOGY AND PRIVACY: THE NEW LANDSCAPE* 193, 198-99 (Philip E. Agre & Marc Rotenberg eds., 1997).

71. See *Dep't of the Air Force v. Rose*, 425 U.S. 352, 382 (1976).

72. *Id.* at 380 n.19.

73. *U.S. Dep't of State v. Washington Post Co.*, 456 U.S. 595, 600-01 (1982).

74. See *U.S. Dep't of Justice v. Reporters Comm. for Freedom of the Press*, 489 U.S. 749, 763-74, 774, 780 (1989).

75. *U.S. Dep't of State v. Ray*, 502 U.S. 164, 176-77 (1991).

76. *U.S. Dep't of Def. v. Fed. Labor Relations Auth.*, 510 U.S. 487, 500-01 (1994) (emphasis omitted from first quotation).

77. See Martin E. Halstuk & Charles N. Davis, *The Public Interest Be Damned: Lower Court Treatment of the Reporters Committee "Central Purpose" Reformulation*, 54 *ADMIN. L. REV.* 983 (2002).

example, a county agency denied a freedom of information request for names and addresses of handgun permit holders citing privacy and safety concerns, but a judge later ordered the county to make the records available.<sup>78</sup> In addition, courts have held that information not easily traced to a particular individual does not constitute an invasion of privacy. For example, the D.C. Circuit held that the Department of the Navy erred in withholding the names and quantities of prescription drugs provided to the Office of Attending Physician to the U.S. Congress because “it is fanciful to assume that without more [information] the knowledge that *someone* among 600 possible recipients was probably using the drug . . . would lead to the conclusion that Beneficiary X has disease Y.”<sup>79</sup> Nevertheless, in some cases, an agency may properly determine that sensitive information could be inferred from a release; for example, disclosing information about individual farmers’ crops and acreage could enable a third party to learn about a farmer’s finances.<sup>80</sup> If a request is drawn narrowly such that the response would unavoidably disclose privacy-sensitive information about an individual or redaction would otherwise not adequately safeguard privacy, an agency may withhold the records, or decline to confirm or deny the existence of any responsive records.<sup>81</sup>

There is evidence that the standards articulated by the judiciary, although they provide support for litigation of FOIA appeals, have very little impact on the release decisions of administrators in practice.<sup>82</sup> Rather, case-by-case determinations regarding the information to withhold or release in response to a FOIA request often vary according to the “position, background, and training” of the official making the decision.<sup>83</sup>

---

78. Jorge Fitz-Gibbon, *Putnam Must Release Gun Records, Judge Says*, THE JOURNAL NEWS (White Plains, N.Y.) (Mar. 5, 2014), <http://www.lohud.com/story/news/2014/03/05/journal-news-putnam-gun-map-lawsuit/6097983>.

79. *Arieff v. U.S. Dep’t of Navy*, 712 F.2d 1462, 1467 (D.C. Cir. 1983) (emphasis in original).

80. *See, e.g., Multi Ag Media LLC v. U.S. Dep’t of Agric.*, 515 F.3d 1224, 1230 (D.C. Cir. 2008).

81. *See Dep’t of the Air Force v. Rose*, 425 U.S. 352, 381 (1976); *see also, Claudio v. Soc. Sec. Admin.*, No. Civ.A. H-98-1911, 2000 WL 33379041, at \*8–9 (S.D. Tex. May 24, 2000) (affirming agency’s decision not to confirm or deny existence of records of investigation of named administrative law judge).

82. *See, e.g., Lillian R. BeVier, Information About Individuals in the Hands of Government: Some Reflections on Mechanisms for Privacy Protection*, 4 WM. & MARY BILL RTS. J. 455, 495 (1995).

83. Lotte E. Feinberg, *Managing the Freedom of Information Act and Federal Information Policy*, 46 PUB. ADMIN. REV. 615, 617 (1986).

## c) Privacy Interventions in Use

In general, agencies protect privacy by withholding or redacting identifiable or sensitive information about individuals. FOIA requires agencies to provide requesters with any reasonably segregable, non-exempt information contained in responsive documents and strongly encourages them to indicate the amount of information redacted from each document, if technically feasible and if doing so would not harm the interest being protected.<sup>84</sup> The types of information commonly redacted include an individual's name, Social Security number, date and place of birth, address, telephone number, criminal history, medical history, and employment history.<sup>85</sup> In some cases, state freedom of information laws similarly prohibit the release of identifiable information such as the names, addresses, and telephone numbers of victims contained within police records.<sup>86</sup> Agencies sometimes take additional steps beyond withholding or redaction to protect data they consider sensitive. For example, when releasing individual-level data about taxi trips, the New York City Taxi Commission attempted to protect taxi drivers' privacy by obscuring all hack license numbers and medallion numbers in the released set of data. However, the commission used a simple hash function that ultimately provided ineffective privacy protection.<sup>87</sup>

The Privacy Act's redress mechanisms are widely considered to be rather weak. To enforce her rights under the Privacy Act, an individual would have to be aware of her rights under the Act, monitor governmental uses and redisclosures of her personal information, identify improper agency actions, and sue the agency in federal court.<sup>88</sup> Even then, the Act limits potential remedies to injunctions requiring an agency to correct the

---

84. See 5 U.S.C. § 552(b).

85. See, e.g., U.S. Dep't of State v. Washington Post Co., 456 U.S. 595, 600 (1982); Associated Press v. U.S. Dep't of Justice, 549 F.3d 62, 65 (2d Cir. 2008).

86. Compare ARK. CODE. ANN. § 16-90-1110(c)(2) (exempting names of victims of crimes and immediate family members from disclosure under freedom of information law), with COLO. REV. STAT. § 24-72-304(4) (2011) (requiring deletion of names and other identifying information about sexual assault victims, but not victims of other crimes, from criminal justice records before release).

87. See Dan Goodin, *Poorly Anonymized Logs Reveal NYC Cab Drivers' Detailed Whereabouts: Botched Attempt to Scrub Data Reveals Driver Details for 173 Million Taxi Trips*, ARS TECHNICA (June 23, 2014, 11:25 AM), <http://www.arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts>; Vijay Pandurangan, *On Taxis and Rainbows—Lessons from NYC's Improperly Anonymized Taxi Logs*, MEDIUM (June 21, 2014), <https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1>.

88. See, e.g., BeVier, *supra* note 82, at 479–82.

individual's record or to produce records wrongly withheld, or actual damages if the individual demonstrates that the agency's intentional or willful violation had an "adverse effect" on her.<sup>89</sup> As Paul Schwartz has argued, this means in practice that "individuals who seek to enforce their rights under the Privacy Act face numerous statutory hurdles, limited damages, and scant chance to effect [sic] an agency's overall behavior."<sup>90</sup> One appeals court, for instance, held that an agency's negligent actions did not violate the law even though the trial court had found that the privacy violations had been "substantial."<sup>91</sup>

Freedom of information laws are a burdensome mechanism for releasing information. Freedom of information decisions are discretionary; the management of requests and compliance is decentralized; and there is little oversight. These realities engender extensive delays, sometimes amounting to years or even decades, that hamper the effectiveness of freedom of information laws.<sup>92</sup> Procedures for requesting and receiving large sets of data are criticized as inefficient. To illustrate, a data analyst who recently sought data about New York City taxi trips was required to purchase and deliver to the taxi commission's offices an unopened 200 GB capacity hard drive and then return to retrieve the hard drive, to which the files had been added, the following day.<sup>93</sup> Agencies are continually experimenting with new ways to make the FOIA process more efficient. Federal agencies are now required to host frequently requested records in electronic reading rooms or libraries.<sup>94</sup> In 2012, the government launched FOIAonline,<sup>95</sup> a web-based tool to help users track the progress of open requests, communicate directly about their status, and access documents that have previously been released. These inefficiencies are also a motivating factor driving the deployment of government open data platforms, discussed below.<sup>96</sup>

---

89. § 552a(g)(2)-(4).

90. Schwartz, *supra* note 7, at 596.

91. *Andrews v. Veterans Admin.*, 838 F.2d 418, 421, 425 (10th Cir. 1988).

92. In 2014, the total backlog of FOIA requests across the federal government was 159,741. U.S. Dept. of Justice, *supra* note 41.

93. See Chris Whong, *FOILing NYC's Taxi Trip Data*, BLOG (Mar. 18, 2014), [http://www.chriswhong.com/open-data/foil\\_nyc\\_taxi](http://www.chriswhong.com/open-data/foil_nyc_taxi) (describing the author's experience with requesting data under New York's Freedom of Information Law).

94. See, e.g., *FOIA Library*, U.S. CENSUS BUREAU, [http://www.census.gov/about/policies/foia/foia\\_library.html](http://www.census.gov/about/policies/foia/foia_library.html) (last modified Oct. 10, 2014).

95. FOIAONLINE, <https://foiaonline.regulations.gov> (last visited May 6, 2015); Nicole Johnson, *Agencies Launch Public FOIA Website*, FEDLINE (Oct. 1, 2012), <http://fedline.federaltimes.com/2012/10/01/agencies-launch-public-foia-website>.

96. See *infra* Section II.A.4.

## 2. *Traditional Public and Vital Records*

State governments have historically made certain records available for inspection as public and vital records. Examples include birth and death certificates, voter registration records, arrest records, civil and criminal court records, bankruptcy filings, professional and business licenses, and property ownership and tax assessment records, among many others. The public availability of these records promotes the transparency of governmental proceedings, actions, and decisions and the facts and rationales underlying these decisions; enables certain transactions such as selling property or initiating lawsuits; and helps individuals learn more about public officials and the people with whom they are considering entering into relationships of trust, such as job candidates or childcare professionals.<sup>97</sup> Public records help members of the public, including journalists, learn about criminal and police activity in their neighborhoods, investigate the prevalence of public safety issues they encounter, and advocate reforms based on the patterns they discover.<sup>98</sup> However, the release of information from public records is sometimes controversial, as evidenced by the public outcry and lawsuits that followed the publication of online maps showing the names, locations, employers, occupations, and contribution amounts of individuals who financially supported a ballot initiative banning same-sex marriage.<sup>99</sup> Journalists and LGBT advocates obtained these records under a state campaign finance disclosure law intended to promote transparency in elections, and then published the records in a way that reportedly led to some harassment and intimidation of donors.<sup>100</sup>

Public records are being made more widely available through increasingly digital and open mechanisms. A significant byproduct is the depreciation of the practical obscurity that once offered some protection to the personal information in these records.<sup>101</sup> Historically, there were practical barriers limiting access to vital and public records, such as the

---

97. See Daniel J. Solove, *Access and Aggregation: Public Records, Privacy, and the Constitution*, 86 MINN. L. REV. 1137, 1173–76 (2002).

98. See *id.*

99. See *ProtectMarriage.com v. Bowen*, 752 F.3d 827, 835 (9th Cir. 2014).

100. See Brad Stone, *Prop 8 Donor Web Site Shows Disclosure Law Is 2-Edged Sword*, N.Y. TIMES (Feb. 7, 2009), <http://www.nytimes.com/2009/02/08/business/08stream.html>.

101. See generally David R. O'Brien et al., *Integrating Approaches to Privacy Across the Research Lifecycle: When Is Information Purely Public?* (Berkman Ctr. for Internet & Soc'y, Research Pub. No. 2015-7, 2015), <http://www.ssrn.com/abstract=2586158> (discussing the gap between expectations of privacy and the increasing public availability of personal information).

necessity of visiting a local office in person during regular business hours to physically search and inspect available records.<sup>102</sup> Locating records of interest through this process could involve trips to multiple offices and significant expenditures of time and money. Alternatively, some agencies have traditionally offered to perform a search and mail relevant records to a requester, assessing a fee for searching for and producing photocopies of the relevant records.

Over time, as these records have been digitized, data management costs have fallen, and data have increasingly been made available online, the barriers to access have diminished significantly for some agencies and types of records. Many public records can now be remotely located through a searchable web-based interface, viewed immediately, and easily linked to information from other sources, though access restrictions vary by court and by jurisdiction. The State of Virginia, for example, makes some records from selected courts available to the public through secure remote access systems, which require prospective users to provide their contact information to the local county clerk's office, pay a \$50 per month subscription fee, and sign an agreement promising not to sell, redistribute, or use the data for improper or illegal purposes.<sup>103</sup> In contrast, the State of Rhode Island makes electronic court records available to the public through courthouse computer terminals, but grants remote access only to attorneys who are admitted to practice in the state, have registered for remote access, and have signed a subscription agreement.<sup>104</sup>

Some state and local agencies will disclose information only in response to targeted requests for individual records, while others will provide information in bulk. In addition, some agencies sell records to commercial information brokers, which in turn manage systems that host the information in fee-based online databases. Private companies, such as data brokers and app developers, are compiling information from public records, combining it with information from other sources, and repackaging the combined information as new products or services.

---

102. See, e.g., CAL. VEH. CODE § 1808(a) (West 2015) (“[A]bstracts of accident reports required to be sent to the [state] . . . shall be open to public inspection during office hours.”).

103. See, e.g., *Remote Access Site*, CITY OF CHESAPEAKE CLERK OF CIRCUIT COURT, <http://www.chesapeakecland.org> (last visited Apr. 28, 2015) (“providing access to land and other related records maintained by this office”).

104. See *Access to Case Information*, RHODE ISLAND JUDICIARY, <https://www.courts.ri.gov/Pages/access-caseinfo.aspx> (last visited July 13, 2015).

LexisNexis, for example, provides a database for mining over 36 billion public records collected from state agencies.<sup>105</sup>

a) Types of Information Released

Depending on the jurisdiction and the type of record, the scope of personal information released in public records may vary. Vital records such as birth, marriage, divorce, and death records often include an individual's name, gender, date and place of birth, and address. Department of motor vehicle records generally include this information plus an individual's Social Security number, disability status, height, weight, eye color, and photograph. Worker's compensation records may also include Social Security numbers, as well as detailed records of the extent of an injury. State employee personnel records may include job titles and salaries. Property ownership and tax assessment records typically contain information, such as size and assessment value, that can reflect the owner's financial situation. Arrest records<sup>106</sup> and sex offender databases may contain names, dates of birth, and photographs, and this information may be made available to the public through a searchable online web interface.<sup>107</sup> Mug shots from police department records are generally deemed to be public records open to inspection, though some jurisdictions exempt them from disclosure or prohibit third parties from misusing the images (e.g., by making it a crime to republish the photographs to a web site that charges subjects of the photos a fee for removal).<sup>108</sup>

---

105. See LEXISNEXIS, TEN COMPELLING REASONS TO RELY ON LEXISNEXIS PUBLIC RECORDS AS YOU RESEARCH PEOPLE, BUSINESSES, AND LOCATIONS (2012), [http://www.lexisnexis.com/pdf/Ten%20Reasons\\_Corp\\_Gov\\_FINAL.pdf](http://www.lexisnexis.com/pdf/Ten%20Reasons_Corp_Gov_FINAL.pdf).

106. See, e.g., IND. CODE § 5-14-3-5(a) (2015) (making available for inspection arrest records including individuals' identifying information such as name, age, and address; charges; and information relating to the circumstances of arrest).

107. See, e.g., *Maine Sex Offender Registry*, MAINE STATE POLICE, <http://sor.informe.org> (last visited Aug. 16, 2015) (provides a full name, date of birth, photograph, town of domicile, place of employment, and list of convictions for sex offender registrants).

108. See, e.g., CAL. CIV. CODE § 1798.91.1(b) (West 2015) ("It shall be unlawful practice for any person engaged in publishing or otherwise disseminating a booking photograph through a print or electronic medium to solicit, require, or accept the payment of a fee or other consideration from a subject individual to remove, correct, modify, or to refrain from publishing or otherwise disseminating that booking photograph."); MINN. STAT. § 13.82(26)(b) (2014) ("Except as otherwise provided . . . , a booking photograph is public data. A law enforcement agency may temporarily withhold access to a booking photograph if the agency determines that access will adversely affect an active investigation.").

## b) Standards for Making Release Decisions

A patchwork of state and local statutes, common law, and administrative practices govern access to and use of vital and public records. State courts determine the scope of information releases, but actual release decisions are made by the individual agencies that maintain the records. Decisions about how different types of records can be accessed by the public, such as whether they can be retrieved in person, by mail, or online, are typically made by agency employees. In evaluating agencies' release decisions, courts balance individuals' right to privacy against the public's right to information. The Supreme Court has held that the public's right to inspect court records is very strong and rooted in "the citizen's desire to keep a watchful eye on the workings of public agencies, and in a newspaper publisher's intention to publish information concerning the operation of government."<sup>109</sup> But a court may properly decide to prohibit access to sensitive personal information contained in its records, based on "a discretion to be exercised in light of the relevant facts and circumstances of the particular case."<sup>110</sup> State and local public records laws arguably provide weak protection for individual privacy,<sup>111</sup> and judicial opinions provide scant guidance for agencies' release decisions.

## c) Privacy Interventions in Use

Practices for restricting disclosures of personal information from public records vary according to jurisdiction and record type. Some states restrict access to personal information by, for example, prohibiting commercial uses such as marketing<sup>112</sup> and requiring individuals seeking public records to pledge not to use the information for solicitation or marketing.<sup>113</sup> Federal law also restricts the disclosure of state public records in a few narrow categories. The Driver's Privacy Protection Act,<sup>114</sup> for example, prohibits state departments of motor vehicles from disclosing personal information from their motor vehicle records, except under limited circumstances such as release to marketers with a subject's consent. Laws

---

109. *Nixon v. Warner Commc'ns, Inc.*, 435 U.S. 589, 598 (1978) (citations omitted).

110. *Id.* at 599.

111. *See, e.g.*, Solove, *supra* note 97, at 1154–72.

112. *See, e.g.*, VA. CODE ANN. § 46.2-208 (2015) (authorizing release of Virginia driver record information for narrowly defined business purposes but providing that "[n]o such information shall be used for solicitation of sales, marketing, or other commercial purposes").

113. *See, e.g.*, CAL. GOV'T CODE § 6254(f)(3) (2014) (prohibiting use of arrest records "directly or indirectly . . . to sell a product or service . . . and the requester shall execute a declaration to that effect under penalty of perjury").

114. Driver's Privacy Protection Act of 1994, 18 U.S.C. § 2721 (2012).

such as the Family Educational Rights and Privacy Act (FERPA)<sup>115</sup> and the Health Insurance Portability and Accountability Act (HIPAA)<sup>116</sup> prohibit the release of certain education and health care records, respectively. Outside of relatively narrow restrictions such as these, public records are generally made available. In light of state data security breach laws and growing complaints from the public and from privacy watchdog groups, government agencies and courts are growing increasingly concerned with protecting the personal information contained in their records, and are exploring new ways to limit public access to sensitive information.

Birth, marriage, and death certificates are typically available only to the person to whom the record pertains, or to certain family members or representatives of that person, for some extended period after the event such as 100 years after birth or 50 years after death. After that period, they typically become publicly available. Depending on the state, voter registration records may be accessible only to political candidates and parties, or may be public and usable for any purpose, including commercial purposes. Federal judges sometimes issue protective orders shielding information from disclosure that might cause an individual “annoyance, embarrassment, oppression, or undue burden or expense.”<sup>117</sup> In particularly sensitive circumstances, a court may determine that a party’s privacy interests outweigh the public’s right to disclosure and seal the records of a proceeding or allow a party to use a pseudonym. In other cases, a court may hold that the public interest in disclosure outweighs the privacy interests. For example, a judge ordered an agency to release citations for violations at state facilities for persons with developmental disabilities because a state law classified the citations as public records.<sup>118</sup> In doing so, it required the records to be released almost in full, subject only to redaction of the names of the individuals receiving services.<sup>119</sup>

Many examples demonstrate the difficulty of making release decisions and adequately safeguarding personal information when subject to state

---

115. Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g (2012).

116. HIPAA Privacy Rule, 45 C.F.R. pt. 160 and subpts. A & E of pt. 164 (2014).

117. FED. R. CIV. P. 26(c).

118. State Dep’t of Pub. Health v. Superior Court, 342 P.3d 1217 (Cal. 2015).

119. *Id.* at 1223 (citing CAL. HEALTH & SAFETY CODE § 1439 (West 2015) (“[T]he names of any persons contained in such records, except the names of duly authorized officers, employees, or agents of the state department conducting an investigation or inspection in response to a complaint filed pursuant to this chapter, shall not be open to public inspection and copies of such records provided for public inspection shall have such names deleted.”)).

public records laws. For instance, in 2003, the county clerk for a Virginia court digitized many of the court's public records to make them available online.<sup>120</sup> When legislators and privacy advocates objected citing the presence of Social Security numbers, dates of birth, and maiden names in the records, the program was suspended so a task force of government attorneys, legislators, privacy experts, and citizens could review and change the system.<sup>121</sup> The county clerk argued that the public records law would have to be amended for him to be able to redact personal information from court records or require individuals to state a permissible purpose before being granted access.<sup>122</sup> He also expressed concern that rejecting an application for access, even one from an individual who had a prior conviction for fraud, could result in a lawsuit for failure to comply with the public records law.<sup>123</sup>

### 3. *Official Statistics*

Designated government agencies prepare and release official statistical information, such as census records and labor statistics, to support policy and business decisions, public transparency, and scientific research.<sup>124</sup> Official statistics are derived from tabular or relational data and measure characteristics of individuals and organizations generated through interviews, questionnaires, and other forms of data collection. Derived official statistics, such as the unemployment index, inform policy analysis and often have legal and regulatory weight in their own right. The Census Bureau, for example, in conducting the decennial census, collects demographic information, such as age, sex, race, and ethnicity, from residents of the United States, supplementing and validating the data collected with administrative records such as tax, Social Security, and municipal records.<sup>125</sup> The statistics it produces are used to draw political districts, apportion seats in the U.S. House of Representatives, distribute

---

120. See Dan Telvock, *Board Passes Resolution to Delay Remote Access of Public Court Records that Contain Personal Data*, LEESBURG TODAY, July 21, 2003.

121. See *id.*

122. See *id.*

123. See *id.*

124. See, e.g., *Databases, Tables & Calculators by Subject*, U.S. BUREAU OF LABOR STATISTICS, <http://www.bls.gov/data> (last visited May 26, 2015).

125. See Lawrence H. Cox & Laura V. Zayatz, *An Agenda for Research in Statistical Disclosure Limitation*, 11 J. OFFICIAL STAT. 205 (1995).

federal funds across the country, and guide the decisions of governments and businesses, among many other uses.<sup>126</sup>

Statistical agencies employ strict confidentiality protections, backed by federal laws such as the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA),<sup>127</sup> to maintain public trust, ensure data integrity, and promote the sustainability of statistical programs.<sup>128</sup> A key privacy threat is the identification of an individual in published data, which is a violation of law and threatens public confidence in statistical agencies' collection and analysis of personal information.<sup>129</sup> Public use data files released by statistical agencies can potentially be linked to other government or commercial data sources, such as voter registration files and social media posts, to uniquely identify individuals.<sup>130</sup> Another threat is inappropriate integration of different types of data across multiple government organizations, which is legally constrained in part by the public's expectations about how the government uses their personal information and general concerns about government surveillance.<sup>131</sup> Commercial firms are also concerned about official statistics leaking their competitive information.<sup>132</sup>

#### a) Types of Information Released

To inform public policy and academic research, statistical agencies release statistical summary data to other agencies and to the general public. The Census Bureau routinely releases data from its surveys and censuses to the public. For example, it releases summary data on population by geographic area, which are used for congressional and state redistricting, as well as summary data on demographic characteristics such

---

126. See U.S. CENSUS BUREAU, MEASURING AMERICA: THE DECENNIAL CENSUSES FROM 1790 TO 2000 (2002), <https://www.census.gov/history/pdf/measuringamerica.pdf>.

127. Confidential Information Protection and Statistical Efficiency Act of 2002, 44 U.S.C. § 3501 note (2012).

128. See OFFICE OF MGMT. & BUDGET, IMPLEMENTATION GUIDANCE FOR TITLE V OF THE E-GOVERNMENT ACT, CONFIDENTIAL INFORMATION PROTECTION AND STATISTICAL EFFICIENCY ACT OF 2002 (CIPSEA) (2006), [https://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/proposed\\_cispea\\_guidance.pdf](https://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/proposed_cispea_guidance.pdf).

129. See *id.*

130. See, e.g., Sweeney, *supra* note 20 (describing a record linkage attack on de-identified health data using public sources).

131. See Stephen E. Fienberg, *Toward a Reconceptualization of Confidentiality Protection in the Context of Linkages with Administrative Records*, 3 J. PRIVACY & CONFIDENTIALITY 65 (2011).

132. See Kinney et al., *supra* note 25.

as age, gender, race, and ethnicity of the total population of the United States.<sup>133</sup> The Bureau of Labor Statistics releases statistics on employment and unemployment rates at the national, state, and local levels; average wages by geographic area and occupation; and average consumer expenditures on food, clothing, and other purchases; among other measures.<sup>134</sup> The National Center for Education Statistics provides statistics on primary and secondary school enrollment by state; graduation and dropout rates; employment of and average salaries for teachers; assessment scores in reading, mathematics, and science by state; rates of college enrollment; and postsecondary degrees awarded.<sup>135</sup>

Agencies disseminate data in various ways, including as derived index data, aggregated tables or sanitized microdata in public use data files, raw data controlled via a secure data enclave, or, to a lesser extent, data made available online through query systems.<sup>136</sup> In some cases, agencies also make available more complex derived tables and, less frequently, geographically aggregated data or sanitized microdata.<sup>137</sup>

#### b) Standards for Making Release Decisions

Producers of official statistics are concerned with a range of disclosures and tend to be highly conservative in releasing data. Laws specifically establish standards for collecting and releasing statistical data. Additionally, based on regulatory requirements, individual agencies have developed specific guidelines for implementing privacy and security safeguards. CIPSEA specifies key standards protecting the confidentiality of data collected by federal agencies for statistical purposes.<sup>138</sup> A primary objective of CIPSEA is to assure survey respondents that their information will not be shared with “regulatory or tax authorities, congressional investigators, prying journalists, or competitors, who might

---

133. See U.S. CENSUS BUREAU: 2010 CENSUS SUMMARY FILE 1: TECHNICAL DOCUMENTATION (2012), <http://www.census.gov/prod/cen2010/doc/sf1.pdf>.

134. See, e.g., *Databases, Tables & Calculators by Subject*, *supra* note 124.

135. See, e.g., THOMAS D. SNYDER & SALLY A. DILLÖW, U.S. DEPT. OF EDUC., DIGEST OF EDUCATION STATISTICS 2013 (2015), <http://nces.ed.gov/pubs2015/2015011.pdf>.

136. See generally LEON WILLENBORG & TON DE WAAL, ELEMENTS OF STATISTICAL DISCLOSURE CONTROL (2001) (discussing in detail the statistical disclosure limitation methodologies used by governments when releasing data).

137. See generally *id.*

138. Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), 44 U.S.C. § 3501 note (2012).

use this information to the detriment of the data provider.”<sup>139</sup> Specifically, CIPSEA protects data collected for statistical purposes by a pledge of confidentiality to the respondent.<sup>140</sup> As required by CIPSEA, statistical agencies review data prior to release to ensure they do not contain information in identifiable form.<sup>141</sup> Many statistical agencies such as the Census Bureau have disclosure review boards, or panels of experts in disclosure limitation, who review each release of summary data, public use data files, statistical estimates or model output, or other information to ensure that it protects confidentiality.<sup>142</sup> For instance, the Census Bureau’s Disclosure Review Board receives data one to two months before the planned date of release, follows a checklist to identify disclosure risks<sup>143</sup> by assessing the statistical disclosure limitation techniques used and the public availability of similar information that could be linked to the data, and recommends techniques for mitigating disclosure risks.<sup>144</sup>

The Privacy Act also exempts the sharing of agency records for statistical research or reporting, as long as the records are “transferred in a form that is not individually identifiable.”<sup>145</sup> Other laws may apply to the statistical activities of particular agencies such as the Internal Revenue Service<sup>146</sup> and the Social Security Administration.<sup>147</sup> For example, Title 13

---

139. Margo Anderson & William Seltzer, *Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues*, 1 J. PRIVACY & CONFIDENTIALITY 7, 8 (2009).

140. A statistical purpose is defined as “the description, estimation, or analysis or the characteristics of groups, without identifying the individuals or organizations that comprise such groups; and includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support [such] purposes.” CIPSEA § 502(9).

141. Information in identifiable form is defined as “any representation of information that permits the identity of the respondent to whom the information applies to be reasonably inferred by either direct or indirect means.” CIPSEA § 502(4).

142. *See, e.g.*, U.S. CENSUS BUREAU, DISCLOSURE REVIEW BOARD (2001), <https://www.census.gov/srd/sdc/wendy.drb.faq.pdf>.

143. Disclosure risk refers to an assessment of the likelihood that an adversary learns the identity or attributes of an individual subject. Note that this term is used more narrowly than privacy risk, as disclosure risks characterize only identifiability while privacy risks encompass the overall additional expected harm from a collection, storage, or release action on the data.

144. *See* U.S. CENSUS BUREAU, *supra* note 142.

145. Privacy Act of 1974, 5 U.S.C. § 552a(b)(5) (2012).

146. 26 U.S.C. § 6108(c) (2012) (“No publication or other disclosure of statistics or other information required or authorized . . . shall in any manner permit the statistics, study, or any information so published, furnished, or otherwise disclosed to be associated with, or otherwise identify, directly or indirectly, a particular taxpayer.”).

147. 42 U.S.C. § 1306(e)(3) (2012) (“[S]uch reports shall not identify individual patients, individual health care practitioners, or other individuals.”).

of the U.S. Code governs the Census Bureau. Title 13 prohibits the agency from releasing “any publication whereby the data furnished by any particular establishment or individual . . . can be identified,”<sup>148</sup> and prohibits the use of statistical data for any purposes other than the statistical purposes for which it was supplied.<sup>149</sup> In addition, all census information protected by Title 13 confidentiality provisions is exempt from disclosure under FOIA. However, Title 13 does not restrict access to or use of census information once it has been publicly released by the Census Bureau.

c) Privacy Interventions in Use

Producers of official statistics employ a number of disclosure limitation methods. Their techniques generally differ for public use data (i.e., data made publicly available without restrictions on access or use) and for restricted use data (i.e., data made available only with strict controls). In general, CIPSEA requires statistical agencies to ensure that the data are handled in a way that minimizes the disclosure risks “throughout the lifecycle of the statistical activity,”<sup>150</sup> that identifiable information is removed before dissemination, and that all employees who have access to the protected data are supervised and controlled. To prepare public use data files, agencies often remove identifiable information prior to publication by using static statistical disclosure controls such as aggregation, suppression, noise addition, and recoding of individual-level data, as well as table-specific suppression and perturbation methods for aggregate data.<sup>151</sup> Common techniques include redacting identifiers, coarsening attributes such as location, recoding values as rounded values or intervals, swapping values in similar records, truncating extreme values, and adding random noise.<sup>152</sup> Agencies make public use datasets available under open access terms without restriction on use or redisclosure. This puts the burden entirely on the agency to mitigate disclosure risks in the public use data files.

---

148. 13 U.S.C. § 9(a)(2) (2012).

149. 13 U.S.C. § 9(a)(1).

150. 72 Fed. Reg. 33362, 33371 (June 15, 2007).

151. *See generally* U.S. CENSUS BUREAU, CENSUS CONFIDENTIALITY AND PRIVACY, 1790–2002 (2003), <http://www.census.gov/prod/2003pubs/conmono2.pdf> (describing Census Bureau confidentiality practices generally); FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16 (providing an overview of statistical disclosure limitation techniques such as perturbation, aggregation, and suppression).

152. *See* FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16.

For restricted use data, researchers generally must apply for access. A formal screening process requires them to provide justification for their request and describe the scope of their research.<sup>153</sup> Some agencies conduct background investigations on prospective researchers and hold them to the same confidentiality standards, backed by criminal penalties, as agency employees.<sup>154</sup> Researchers' use of restricted data is limited to the purposes they specified, and access is restricted to that necessary for the proposed analysis.<sup>155</sup> Data use agreements often bind the researcher to specific use and disclosure restrictions, and violations of confidentiality provisions may carry significant legal or even criminal penalties.<sup>156</sup> Agencies also employ technical controls on access and use via research data centers or enclaves<sup>157</sup> or, less frequently, remote analysis servers, which allow access to dynamically derived tables and maps.<sup>158</sup> Some large statistical agencies are also experimenting with emerging computational techniques such as synthetic data and differential privacy. For example, the Census Bureau has produced a tool called OnTheMap, which implements a variant of differential privacy to map workforce related data in a privacy-preserving way.<sup>159</sup> Statistical agencies often evaluate the effectiveness of their disclosure limitation techniques by performing privacy impact assessments and staging reidentification attacks using available auxiliary datasets. For data users whose needs are not met by the public use data files, an agency may have a program for generating custom tabulations which are screened by a disclosure review board before release.<sup>160</sup>

---

153. See, e.g., U.S. CENSUS BUREAU, CENSUS RDC RESEARCH PROPOSAL GUIDELINES 1–12 (2015), [https://www.census.gov/ces/pdf/Research\\_Proposal\\_Guidelines.pdf](https://www.census.gov/ces/pdf/Research_Proposal_Guidelines.pdf) (describing the process of applying for access to research data through the Census Bureau Research Data Center).

154. See, e.g., *id.* at 12.

155. See, e.g., *id.* at 6–10 (describing how to state research purposes); *id.* at 3–4 (explaining that applications must thoroughly address what datasets are being requested and why, and noting that certain sensitive personal information is categorically unavailable at the agency's data centers).

156. See, e.g., *id.*

157. See, e.g., *Federal Statistical Research Data Centers*, U.S. CENSUS BUREAU, <http://www.census.gov/about/adrm/fsrdc/locations.html> (last visited May 28, 2015).

158. See, e.g., Michael Freiman et al., *The Microdata Analysis System at the U.S. Census Bureau*, 2011 JOINT STAT. MEETINGS 3645 (2011) (discussing a Census Bureau remote analysis server currently in development and providing an overview of similar systems that have been proposed or implemented).

159. *OnTheMap*, U.S. CENSUS BUREAU, <http://onthemap.ces.census.gov> (last visited May 28, 2015).

160. See, e.g., *Special Tabulations Program*, U.S. CENSUS BUREAU, <https://www.census.gov/population/www/cen2000/sptabs/main.html> (last visited May 28, 2015).

Emerging challenges in this area include the rising speed of data collection and processing (sometimes referred to as data velocity),<sup>161</sup> heightened data integration,<sup>162</sup> and increasing analytic sophistication.<sup>163</sup> Capabilities for linking statistical data to auxiliary data sources are improving, and common techniques for limiting disclosure risks can greatly diminish the utility of the data.<sup>164</sup> Agencies are pressured to release data faster, more cost effectively, and in a way that allows a greater range of analysis, including visualizations and data mining, and provides estimates for finer time scales and geographic areas.<sup>165</sup>

#### 4. *E-Government and Open Government Initiatives*

Many governments have recently begun implementing e-government and open government initiatives that operate on a “presumption of openness.”<sup>166</sup> In light of technological advances and increasing public demands for data, governments now encourage agencies to “publish information online in an open format that can be retrieved, downloaded, indexed, and searched by commonly used web search applications.” Additionally, governments now encourage agencies to “proactively use modern technology to disseminate useful information, rather than waiting for specific requests under FOIA.”<sup>167</sup> Government agencies at all levels are launching open data repositories, analysis tools, and discussion forums, for viewing, manipulating, downloading, and discussing large quantities of government data. Thus, e-government and open data programs represent a fundamental shift in the way governments release data.

---

161. See EXEC. OFFICE OF THE PRESIDENT, *BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES* (2014), [https://www.whitehouse.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf).

162. See Gerald W. Gates, *How Uncertainty about Privacy and Confidentiality Is Hampering Efforts to More Effectively Use Administrative Records in Producing U.S. National Statistics*, 3 J. PRIVACY & CONFIDENTIALITY 3 (2011); Fienberg, *supra* note 131.

163. See Christian Reimsbach-Kounatze, *The Proliferation of “Big Data” and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis* (OECD Digital Economy Paper No. 245, 2015), [http://www.oecd-ilibrary.org/science-and-technology/oecd-digital-economy-papers\\_20716826](http://www.oecd-ilibrary.org/science-and-technology/oecd-digital-economy-papers_20716826).

164. See, e.g., Carl Bialik, *Census Bureau Obscured Personal Data—Too Well, Some Say*, WALL ST. J. (Feb. 6, 2010), <http://www.wsj.com/articles/SB10001424052748704533204575047241321811712>.

165. See WILLIAM E. WINKLER, U.S. CENSUS BUREAU, RESEARCH REPORT NO. RR98/02, *PRODUCING PUBLIC-USE MICRODATA THAT ARE ANALYTICALLY VALID AND CONFIDENTIAL* (1998), <https://www.census.gov/srd/papers/pdf/rr9802.pdf>.

166. ORSZAG, *supra* note 8; see, e.g., NYC OPEN DATA, <https://nycopendata.socrata.com> (last visited June 29, 2015).

167. ORSZAG, *supra* note 8.

In 2002, the federal government announced an E-Government Strategy aimed at improving the transparency, effectiveness, and responsiveness of governmental services by leveraging digital storage, computing power, Internet connectivity, and related advances of the information age.<sup>168</sup> Its principal aims were to create a “citizen-centered E-Government” that utilizes web services to improve citizens’ interactions with the federal government, and to make recordkeeping more efficient by digitizing and coordinating information collection and storage across agencies and departments.<sup>169</sup>

Building on the e-government efforts, President Obama issued the Open Government Directive in 2009, which ordered all federal executive agencies to make available online as many nonclassified datasets as possible.<sup>170</sup> Specifically, the directive required all agencies to publish at least three previously non-public datasets containing high-value information to further agency accountability and responsiveness, enhance public knowledge, further agency core missions, and create economic opportunity.<sup>171</sup> It also mandated that agencies identify additional high-value information and prepare a timeline for publishing such information online in open formats.<sup>172</sup> In 2011, the Obama administration implemented the Open Government National Action Plan for developing new online tools to increase civic participation, update record management practices, make information from FOIA requests available online, increase declassification of national security information, and improve the implementation of open government plans across agencies.<sup>173</sup>

Most recently, in 2013, President Obama signed an executive order directing the implementation of an Open Data Policy across the federal government requiring that “the default state of new and modernized Government information resources shall be open and machine readable” and that these information resources “shall be managed as an asset throughout its life cycle to promote interoperability and openness and, wherever possible and legally permissible, to ensure that data are released

---

168. OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, E-GOVERNMENT STRATEGY (2002), <https://www.whitehouse.gov/sites/default/files/omb/inforeg/egovstrategy.pdf>.

169. *Id.* at 1–2.

170. ORSZAG, *supra* note 8.

171. *Id.* at 7–8.

172. *Id.*

173. THE OPEN GOVERNMENT PARTNERSHIP, NATIONAL ACTION PLAN FOR THE UNITED STATES OF AMERICA (2011), [https://www.whitehouse.gov/sites/default/files/us\\_national\\_action\\_plan\\_final\\_2.pdf](https://www.whitehouse.gov/sites/default/files/us_national_action_plan_final_2.pdf).

to the public in ways that make the data easy to find, accessible, and usable.”<sup>174</sup> The executive order also requires agencies to “safeguard individual privacy, confidentiality, and national security” when implementing the policy.<sup>175</sup>

a) Types of Information Released

The public release of data plays an essential role in these initiatives, and data that have already been released by open government and e-government programs are extensive and wide ranging. They include communications, representations of knowledge, facts, data, and opinions presented in various mediums and formats. For example, data are offered in static datasets or in real-time streams, provided as tabular data or through data visualization tools, and contain data types such as textual, multimedia, sensor, or geospatial data. Federal, state, and local agencies are making large datasets available online in formats that are free, available for use on a variety of platforms, and open to the public, without restrictions. Journalists, civic groups, researchers, and citizens are now able to reuse data in new ways that promote transparency and accountability, improve the effectiveness and responsiveness of government agencies, and create economic benefits.

Open data are also advancing the state of research and scientific knowledge. Social scientists are increasingly obtaining data from government records, government organizations, businesses such as telephone and utility providers, and sensors such as public thermal imaging cameras. For example, the Boston Area Research Initiative seeks to promote original research by combining new social science model-based approaches, data mining, and other big data methods that combine data from traditional sources with sensor data.<sup>176</sup> The Center for Urban Science and Progress at New York University also uses big data methods and combinations of sensor data (such as thermal imaging) and administrative data to guide urban policymaking and operations.<sup>177</sup> In addition, making

---

174. Exec. Order No. 13,642, 3 C.F.R. 244 (2014) (Making Open and Machine Readable the New Default for Government Information), <https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->.

175. *Id.*

176. See Daniel Tumminelli O'Brien, Robert J. Sampson & Christopher Winship, *Ecometrics in the Age of Big Data: Measuring and Assessing “Broken Windows” Using Administrative Records* (Bos. Area Research Initiative, Working Paper No. 3, 2013).

177. See STEVEN E. KOONIN, CTR. FOR URBAN SCI. & PROGRESS, *THE PROMISE OF URBAN INFORMATICS* (2013), <http://cusp.nyu.edu/wp-content/uploads/2013/07/CUSP-overview-May-30-2013.pdf>.

rich data sources available for free is one way that states and municipalities can attract technology companies to an area and bolster their local economies.<sup>178</sup> Third party data analysts and commercial firms use the data released by government agencies to produce apps such as up-to-the-minute public transit tracking apps.<sup>179</sup>

Increasingly, governments are releasing data through online data portals such as Data.gov, which the Obama administration launched in 2009 as the federal government's central clearinghouse for open data. Agencies proactively post data in raw, structured formats via Data.gov, and these data may be downloaded for free and without any restriction on future use.<sup>180</sup> As of May 2015, 83 agencies and sub-agencies have published over 130,000 datasets on Data.gov<sup>181</sup> (though some point out that many of the datasets were published by just a handful of agencies or are duplicates of datasets previously posted elsewhere online).<sup>182</sup> In 2013, the Obama administration launched Project Open Data, an open source project for implementing open data repositories and related tools for sharing, converting, visualizing, and using data.<sup>183</sup> Project Open Data and similar projects are making data increasingly available through application programming interfaces (APIs), and their APIs give third party software developers direct access to data in formats that can be fed into consumer apps for smartphones and web sites, and, in some cases, enable use and analysis of real-time data streams.<sup>184</sup>

State, county, and local governments are also implementing open data initiatives based on the federal government's model. As of May 2015, thirty-nine states and forty-six cities and counties have launched open data

---

178. See STEPHEN GOLDSMITH & SUSAN CRAWFORD, *THE RESPONSIVE CITY: ENGAGING COMMUNITIES THROUGH DATA-SMART GOVERNANCE* 78–79 (2014).

179. See *Open Data Is Making Transit Better, One App at a Time*, TRIMET BLOG (July 23, 2014), <http://howweroll.trimet.org/2014/07/23/open-data-is-making-transit-better-one-app-at-a-time>.

180. DATA.GOV, <http://www.data.gov> (last visited May 6, 2015).

181. *Id.*

182. See Alon Peled, *When Transparency and Collaboration Collide: The USA Open Data Program*, 62 J. AM. SOC. FOR INFO. SCI. & TECH. 2085, 2088 (2011).

183. See Todd Park & Steven VanRoekel, *Introducing: Project Open Data*, WHITE HOUSE OFF. SCI. & TECH. POL'Y BLOG (May 16, 2013), <https://www.whitehouse.gov/blog/2013/05/16/introducing-project-open-data>.

184. See, e.g., *PHL API*, CITY OF PHILA., <http://www.phlapi.com> (last visited Aug. 16, 2015) (providing open data APIs for property values, polling locations, licenses and permits, 311 reports, crime incidents, geospatial information, and airport parking availability, among other data from the City of Philadelphia).

portals.<sup>185</sup> These open data portals rely on state public records laws to obtain and publish business license records, crime incident reports, 311 service requests, building permits, property assessments, restaurant inspections, and more. Municipal open data can enable analyses integrating large quantities of data from many existing observational sources. Inspired by the public availability of open data, third party developers are creating applications that combine data from multiple sources in ways that create value for the public.<sup>186</sup> For example, RentCheck uses municipal open data to generate a searchable, interactive map with which people can review 311 complaints and inspection violations filed for specific apartment buildings throughout New York City.<sup>187</sup>

At the same time, the release of this information as open data has privacy implications. Much of the data contains information about individuals and may lead to privacy violations. In some cases, government agencies routinely release data in selected formats or to selected parties, but the data are then treated as public in subsequent redisclosures and in linking with other data in ways agencies may not have anticipated. Although such data are often considered public records, the agencies are required to make a determination of the effects of a disclosure on individual privacy. Privacy issues may also emerge with new data sources. For example, sensor data collected in public places nevertheless may include activity occurring on private property, as in the case of sensors that monitor light and pollutants emitted from private buildings.<sup>188</sup> Data are also frequently released with the understanding, which is often documented, that the data have already undergone limited de-identification. However, on receipt, it is sometimes obvious upon reasonable inspection that the data still contains direct or indirect identifiers that may reveal sensitive information about individuals, as described in detail in Section IV.B below.

---

185. See *Open Government*, DATA.GOV, <https://www.data.gov/open-gov> (last visited May 26, 2015).

186. See GOLDSMITH & CRAWFORD, *supra* note 178, at 78.

187. See Karen Eng, *Check before you rent: How a TED Fellow is holding New York City landlords accountable*, TEDBLOG (Apr. 10, 2015), <http://blog.ted.com/how-ted-fellow-yale-fox-is-holding-new-york-city-landlords-accountable>.

188. See Elizabeth Dwoskin, *They're Tracking When You Turn Off the Lights: Municipal Sensor Networks Measure Everything from Air Pollution to Pedestrian Traffic; Building a "Fitbit for the City,"* WALL ST. J. (Oct. 20, 2014), <http://www.wsj.com/articles/theyre-tracking-when-you-turn-off-the-lights-1413854422>.

## b) Standards for Making Release Decisions

While a substantial portion of the data held by government agencies and being considered for release as open data do not directly relate to human characteristics or behaviors (e.g., meteorological or agricultural information), much of the data is related to individuals. Therefore, when collecting, storing, and sharing data about individuals, federal executive agencies must follow certain data security practices prescribed by the National Institute of Standards and Technology (NIST),<sup>189</sup> disclosure limitation practices outlined in the 2005 Federal Committee on Statistical Methodology report,<sup>190</sup> and information privacy provisions in laws such as the Privacy Act of 1974,<sup>191</sup> the E-Government Act of 2002 (including CIPSEA),<sup>192</sup> and the Federal Information Security Management Act.<sup>193</sup> The Open Government Directive, recognizing that there may be privacy risks associated with data slated for release, exempts privacy-sensitive information from release, providing that “[w]ith respect to information, the presumption shall be in favor of openness (to the extent permitted by law and subject to valid privacy, confidentiality, security, or other restrictions.)”<sup>194</sup> Furthermore, the Open Data Policy requires agencies to “incorporate privacy analyses into each stage of the information’s life cycle,” to “review the information collected or created for valid restrictions to release to determine whether it can be made publicly available,” and to work with their “Senior Agency Official for Privacy (SAOP) or other relevant officials to ensure that privacy and confidentiality are fully protected.”<sup>195</sup> The Open Data Policy instructs agencies to conduct a risk-based analysis when deciding whether to release certain information, “often utilizing statistical methods whose parameters can change over

---

189. See, e.g., NIST, FIPS PUB. 199, STANDARDS FOR SECURITY CATEGORIZATION OF FEDERAL INFORMATION AND INFORMATION SYSTEMS (2004) [hereinafter NIST, STANDARDS], <http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf>; NIST, SPECIAL PUB. 800-53, REVISION 4, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS (2013) [hereinafter NIST, CONTROLS (draft)], <http://csrc.nist.gov/publications/drafts/800-53-rev4/sp800-53-rev4-ipd.pdf>.

190. FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16.

191. 5 U.S.C. § 552a.

192. E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899.

193. Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–3549 (2012).

194. ORSZAG, *supra* note 8.

195. OFFICE OF MGMT. & BUDGET, OPEN DATA POLICY—MANAGING INFORMATION AS AN ASSET, MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES (2013), <https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>.

time, depending on the nature of the information, the availability of other information, and the technology in place that could facilitate the process of identification.”<sup>196</sup> Given the complexity of this analysis, agencies “may choose to take advantage of entities in the Executive Branch that may have relevant expertise, including the staff of Data.gov.”<sup>197</sup>

The Open Data Policy also instructs federal agencies to create a public inventory of all data that are or could be made public and assign an access level to each set of data based on a three-tier scheme for controlled unclassified information.<sup>198</sup> In this system, the “public” level permits data to be made publicly available to anyone without restriction, while the “restricted public” level denotes certain use restrictions. An example provided for the “restricted public” classification is data “that can only be made available to select researchers under certain conditions, because the data asset contains sufficient granularity or linkages that make it possible to reidentify individuals, even though the data asset is stripped of Personally Identifiable Information (PII).” Another example is data “that contains PII and is made available to select researchers under strong legal protections.”<sup>199</sup> The third level, “non-public,” is used for data that cannot be made available to the public and may only be shared within the federal government.<sup>200</sup>

At the state and local levels, standards for releasing open data vary widely depending on the jurisdiction, government department, and type of data. As noted above in the discussion of state public and vital records, state laws designate records as public records using different standards, and because open data release decisions rely in large part on state public records laws, there is significant variation in release decisions across state and local open data programs. When granted wide discretion in making release decisions, government departments within the same jurisdiction also develop different standards for releasing open data. Some departments, for instance, are known for making more conservative data sharing decisions for reasons related to the organization’s historical practices, expertise, and interpretation of regulatory obligations. Commentators have observed that department staff often express

---

196. *Id.* at 9–10.

197. *Id.* at 10.

198. PROJECT OPEN DATA, IMPLEMENTATION GUIDE: SUPPLEMENTAL GUIDANCE ON THE IMPLEMENTATION OF M-13-13 “OPEN DATA POLICY — MANAGING INFORMATION AS AN ASSET,” <https://project-open-data.cio.gov/implementation-guide> (last visited May 19, 2015).

199. *Id.*

200. *Id.*

uncertainty regarding regulatory requirements and that government lawyers frequently overinterpret legal standards.<sup>201</sup> For example, government employees express concerns that privacy laws protect data held by their departments, but they lack specific guidance for screening individual datasets prior to release. Due to the existence of a specific privacy law, government actors might also decline to release all data related in a specially regulated space, such as education, due to the existence of FERPA, an information privacy law that protects certain education records.<sup>202</sup> The lack of formal guidance and definitions for determining which datasets, and which fields within the datasets, can be released as open data has led to conflicting opinions between city departments that generate and release datasets. It has also led to a data review process that is time intensive and arguably not sustainable over the long term as vastly more data are set to be released.

c) Privacy Interventions in Use

To assist agencies in systematically reviewing data prior to release and selecting appropriate controls for mitigating disclosure risks, an interagency working group led by the National Security Staff developed more specific guidance for conducting data privacy and security reviews.<sup>203</sup> This guidance expressly recognizes the cumulative “mosaic effect” of releasing pieces of information over time and aims to reduce potential record linkages between a released set of data and other available information.<sup>204</sup> Its central component is a checklist for assessing the privacy risks in datasets submitted for publication to Data.gov. This checklist is completed through an online assessment tool or by filling in a metadata template that accompanies the dataset when it is submitted for publication.<sup>205</sup> The checklist asks whether the dataset has previously undergone a formal disclosure committee review,<sup>206</sup> whether the data were collected from respondents under a promise of confidentiality, and whether a FOIA exemption applies to the information.<sup>207</sup> If the dataset contains microdata (individual-level rather than aggregate information), the checklist asks whether the microdata include direct identifiers

---

201. GOLDSMITH & CRAWFORD, *supra* note 178, at 164–65.

202. *See id.*

203. DATA.GOV, NATIONAL/HOMELAND SECURITY AND PRIVACY/CONFIDENTIALITY CHECKLIST AND GUIDANCE, [http://www.data.gov/sites/default/files/attachments/Privacy and Security Checklist.pdf](http://www.data.gov/sites/default/files/attachments/Privacy%20and%20Security%20Checklist.pdf) (last viewed June 29, 2015).

204. *Id.* at 1–2.

205. *Id.* at 2.

206. *Id.* at 6.

207. *Id.* at 7.

(“information that exclusively identifies a person or business”) or indirect identifiers (“information that, when used in combination with other data, could lead to the identification of a person or business”).<sup>208</sup> The checklist also asks whether any disclosure limitation techniques, such as suppression, top or bottom coding, data swapping, collapsing categories, or data blurring, have been applied to the dataset.<sup>209</sup> Open government data are typically de-identified by redacting direct or indirect identifiers, or applying statistical disclosure limitation techniques. Examples of commonly removed direct identifiers include names, Social Security numbers, dates of birth, addresses, telephone numbers, email addresses, and web universal resource locators (URLs).<sup>210</sup> Indirect identifiers typically include dates other than birthdates, locations and geographic information, and demographic characteristics such as gender or age.<sup>211</sup>

Agencies have established disclosure review practices for releases of information to the public, and the working group guidance and checklist described above have supplemented but not replaced these practices. Other established agency practices for reviewing data prior to release include performing privacy impact assessments and assigning general access levels to data based on guidance from the Office of Management and Budget,<sup>212</sup> NIST,<sup>213</sup> and Controlled Unclassified Information program<sup>214</sup> documentation.<sup>215</sup> For instance, the E-Government Act of 2002<sup>216</sup> requires federal executive agencies to perform privacy impact assessments for their electronic information systems and any identifiable information about individuals they contain.<sup>217</sup> The Act directs agencies completing these assessments to examine the privacy risks and effects of collecting, storing, and disseminating identifiable information about individuals, to describe how electronic information will be handled in accordance with legal, regulatory, and policy requirements for privacy; and to specify the practices that will be put in place to mitigate privacy risks.<sup>218</sup>

---

208. *Id.* at 8.

209. *Id.* at 10.

210. *Id.* at 12.

211. *Id.* at 13–14.

212. OFFICE OF MGMT. & BUDGET, *supra* note 195.

213. NIST, STANDARDS, *supra* note 189.

214. Exec. Order No. 13,556, 3 C.F.R. 267 (2011).

215. *See* PROJECT OPEN DATA, *supra* note 198.

216. Pub. L. No. 107-347, 116 Stat. 2899

217. OFFICE OF MGMT. & BUDGET, MEMORANDUM M-03-22, OMB GUIDANCE FOR IMPLEMENTING THE PRIVACY PROVISIONS OF THE E-GOVERNMENT ACT OF 2002 (2003), [https://www.whitehouse.gov/omb/memoranda\\_m03-22](https://www.whitehouse.gov/omb/memoranda_m03-22).

218. *Id.*

Factors covered in a privacy impact assessment include the nature and source of information to be collected, the purpose for the collection, the intended use of the information, the intended recipients of the information, the opportunities to consent or decline to provide information, the information security controls, and whether the Privacy Act would apply.<sup>219</sup> The Act also requires agencies to “consider the information ‘life cycle’ (i.e., collection, use, retention, processing, disclosure and destruction) in evaluating how information handling practices at each stage may affect individuals’ privacy” and to consult “program experts as well as experts in the areas of information technology, IT security, records management and privacy” in these assessments.<sup>220</sup>

#### B. SHORTCOMINGS IN CURRENT PRACTICES

The foregoing discussion of many common approaches to releasing government data reflects wide variation in scope, sources, purpose, and regulatory constraints across use cases. It also reveals three potential shortcomings related to the protection of individual privacy in such releases. This Section identifies these commonly occurring shortcomings in privacy analysis and protection within the broad categories of data releases. In Part III, we argue that these observations demonstrate the need for a more comprehensive framework for characterizing and aligning the utility, threats, vulnerabilities, and controls associated with a given data release.

The first shortcoming is that, in contrast to the wide variety of scenarios within which governments release data, the approach that most government actors take with respect to privacy is rather narrow and homogenous. Despite differences in regulatory language and context, most agencies, with the notable exception of large statistical agencies, address regulatory requirements for privacy protection in the same fashion: by withholding or redacting records that contain certain pieces of directly or indirectly identifying information. For instance, federal agencies releasing information in response to FOIA requests typically remove an individual’s name, Social Security number, date and place of birth, address, telephone number, and information related to medical, employment, or criminal history.<sup>221</sup> Most state agencies similarly protect privacy by withholding categories of records, such as juvenile court records, or identifiable

---

219. *Id.*

220. *Id.*

221. *See, e.g.*, U.S. Dep’t of State v. Washington Post Co., 456 U.S. 595, 600 (1982); Associated Press v. U.S. Dep’t of Justice, 549 F.3d 62, 65 (2d Cir. 2008); *see also* discussion *supra* Section II.A.1.

information in records, such as the names of sexual assault victims in police records, that are deemed to be sensitive.<sup>222</sup> Following standards from state public records laws, municipal open data portals also redact identifiers from datasets before their release, and withhold entirely datasets deemed to be especially sensitive or regulated by a specific information privacy law.<sup>223</sup>

This focus on a small set of controls appears suboptimal. It is now a well-established principle in the scientific literature that privacy risks are not a simple function of the presence or absence of specific fields, attributes, or keywords in a released set of data.<sup>224</sup> Instead, much of the potential for harm stems from what one can infer about individuals from the data release as a whole or when the data are linked with other available information. It generally takes very little information to uniquely identify an individual.<sup>225</sup> There have been numerous examples where this phenomenon has been exploited for reidentification, even with seemingly innocuous information that falls outside the scope of what is considered to be directly or indirectly identifying information.<sup>226</sup> Government releases of information that involve an ad-hoc balancing of interests or redactions of certain fields will likely fail to address the nuances of privacy risks. As a result, governments that rely only on redaction likely disclose information that exposes individuals to privacy risks or withhold useful information that could be safely shared.

The second shortcoming is that guidance on interpreting and applying regulatory standards for privacy protection appears remarkably thin. In recent draft guidelines, NIST noted that “[a]lthough existing tools such as the Fair Information Practice Principles (FIPPs) and privacy impact assessments (PIAs) provide a foundation for taking privacy into consideration, they have not yet provided a method for federal agencies to measure privacy impacts on a consistent and repeatable basis.”<sup>227</sup> General

---

222. See discussion *supra* Section II.A.2.

223. See discussion *infra* Section IV.B.

224. See, e.g., Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, PROCEEDINGS OF THE 2008 IEEE SYMPOSIUM ON RESEARCH IN SECURITY AND PRIVACY 111 (2008); Latanya Sweeney, *k-anonymity: A Model for Protecting Privacy*, 10 INT’L J. OF UNCERTAINTY FUZZINESS & KNOWLEDGE-BASED SYSTEMS 557 (2002).

225. See, e.g., Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCIENCE 536 (2015).

226. See *id.*

227. NIST, PRIVACY RISK MANAGEMENT FOR FEDERAL INFORMATION SYSTEMS, NISTIR 8062 (DRAFT) 1 (2015), [http://csrc.nist.gov/publications/drafts/nistir-8062/nistir\\_8062\\_draft.pdf](http://csrc.nist.gov/publications/drafts/nistir-8062/nistir_8062_draft.pdf).

guidance directing agencies to protect the privacy of individuals and prevent the release of personally identifiable information is common, yet there is relatively little regulatory guidance for formally characterizing privacy risks and selecting and implementing controls and interventions in specific settings. The literature review, use case analysis, and expert interviews used for the case studies in this Article reveal only a handful of well-recognized or widely adopted sources on identifying and mitigating privacy risks.<sup>228</sup> In addition, on the whole this formal guidance is general, abstract, infrequently updated, and self-directed.<sup>229</sup> Guidelines for implementing the formal guidance within specific agencies, legal frameworks, and data releases are essential, yet agencies typically point to these materials without providing direction for their implementation.<sup>230</sup> In contrast, formal guidance for analyzing and mitigating related information security risks, such as that described in FISMA,<sup>231</sup> is voluminous, proscriptive, specific, actionable, frequently updated, and integrative into legal systems of audit and certification.<sup>232</sup> The comparative paucity of

---

228. See, e.g., FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16; NIST, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS, SPECIAL PUB. 800-53 (2013) [hereinafter NIST, CONTROLS (final)], <http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf>; NIST, GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII), SPECIAL PUB. 800-122 (2010), <http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf>; OFFICE OF MGMT. & BUDGET, MEMORANDUM RE: SAFEGUARDING AGAINST AND RESPONDING TO THE BREACH OF PERSONALLY IDENTIFIABLE INFORMATION (May 22, 2007), <https://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf>; U.S. DEP'T OF HEALTH, EDUC., & WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS: REPORT OF THE SECRETARY'S ADVISORY COMMITTEE ON AUTOMATED PERSONAL DATA SYSTEMS (1973), <http://www.justice.gov/opcl/docs/rec-com-rights.pdf>.

229. For example, one of the most frequently cited guidance documents on this subject, the Federal Committee on Statistical Methodology Report on Statistical Disclosure Limitation Methodology, was last revised in 2005. FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16. In addition, the report provides an introduction to statistical concepts and techniques for disclosure limitation, but it does not provide direction on selecting among the available techniques for application in a particular data release. See *id.*

230. See, e.g., U.S. Department of Justice, Office of Privacy and Civil Liberties: Resources, <http://www.justice.gov/opcl/resources> (last visited June 1, 2015).

231. Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–49 (2012).

232. See discussion *infra* Section III.B.

privacy documentation often leads to inconsistent identification of privacy risks and ineffective application of privacy safeguards.<sup>233</sup>

The third shortcoming is that similar privacy risks—and, in some cases, even identical data—are treated quite differently by different government actors. This is most apparent in the ways in which governments evaluate the source and degree of privacy risk. Depending on the context, government releases of information are subject either to laws and regulations that protect privacy by requiring a balancing of interests for and against disclosure, or to laws and regulations that protect privacy by prohibiting the release of any information deemed to be personally identifiable. FOIA, for example, falls into the first category, as it compels agencies to release information to the public, but grants them discretion to withhold certain types of information that “would constitute a clearly unwarranted invasion of personal privacy” if released.<sup>234</sup> Examples in the latter category include state freedom of information laws that expressly require redaction of identifying information about sexual assault victims.<sup>235</sup> The Privacy Act similarly prohibits the release of information such as an individual’s education, financial, medical, criminal, or employment history in combination with a name or “other identifying particular assigned to the individual,”<sup>236</sup> and statistical agencies are likewise prohibited from disclosing information about individuals in identifiable form.<sup>237</sup>

In some cases, the same measurements of the same people are provided with different protections as the data move from agency to agency. For example, because CIPSEA governs the Bureau of Labor Statistics, the Bureau releases only aggregate statistics based on information collected from Occupational Safety and Health Administration (OSHA) logs, even though OSHA is permitted to release establishment-level and individual-level records from the same logs.<sup>238</sup> These observations suggest that release decisions and the use of privacy controls are not well matched to the privacy risks associated with a specific set of data.

---

233. For an in-depth discussion of some of the challenges and gaps that data managers have encountered in interpreting and applying general regulatory guidance in specific data release cases, see discussion *infra* Section IV.B.

234. 5 U.S.C. § 552(b)(6).

235. *See, e.g.*, COLO. REV. STAT. § 24-72-304(4) (2011).

236. 5 U.S.C. §§ 552a(a)(4), (b).

237. *See* Confidential Information Protection and Statistical Efficiency Act of 2002 § 512(b), 44 U.S.C. § 3501 note (2012).

238. *See* Proposed Rule, Improve Tracking of Workplace Injuries and Illnesses, 78 Fed. Reg. 67254, 67257–60 (Nov. 8, 2013).

### III. A FRAMEWORK FOR MODERNIZING PRIVACY ANALYSIS

As Part II highlights, when governments attempt to manage confidentiality in data releases, they appear to rely on only a few tools and little formal guidance. This results in data releases that are both less useful and less protective than they could be and treatment of data across government actors that is inconsistent.

Governments use a narrow set of tools to analyze and mitigate privacy risks, despite the broad range of privacy interventions proposed by privacy scholars, legal scholars, non-profit organizations, and many others. There exists a broad range of proposals for privacy intervention, operating at different conceptual levels. For example, article 12 of the UN Declaration of Human Rights<sup>239</sup> and the framework of privacy by design<sup>240</sup> contain high-level privacy principles. Fair information practice principles<sup>241</sup> and contextual integrity<sup>242</sup> provide mid-level guidance. Privacy impact assessments,<sup>243</sup>  $k$ -anonymity,<sup>244</sup> and traditional statistical disclosure limitation techniques<sup>245</sup> are examples of applied methods for enhancing confidentiality. Proposals such as differential privacy<sup>246</sup> incorporate formal mathematical frameworks for privacy. Finally, some proposals at the

---

239. G.A. Res. 217 (III) A, Universal Declaration of Human Rights, U.N. Doc. A/RES/217(III), art. 12 (Dec. 10, 1948) (“No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.”).

240. ANN CAVOUKIAN, PRIVACY BY DESIGN 1 (2011), <https://www.ipc.on.ca/images/Resources/7foundationalprinciples.pdf>.

241. See ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, THE OECD PRIVACY FRAMEWORK (2013), [http://www.oecd.org/sti/ieconomy/oecd\\_privacy\\_framework.pdf](http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf); FEDERAL TRADE COMMISSION, PRIVACY ONLINE: FAIR INFORMATION PRACTICES IN THE ELECTRONIC MARKETPLACE: A REPORT TO CONGRESS (2000), <https://www.ftc.gov/sites/default/files/documents/reports/privacy-online-fair-information-practices-electronic-marketplace-federal-trade-commission-report/privacy2000.pdf>; U.S. DEP’T OF HEALTH, EDUC. & WELFARE, *supra* note 228.

242. See Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119 (2004).

243. See DAVID WRIGHT & PAUL DE HERT, PRIVACY IMPACT ASSESSMENT (2012).

244. The  $k$ -anonymity model describes a release in which each record cannot be distinguished from at least  $k-1$  other records. See Sweeney, *supra* note 224.

245. See Gregory J. Matthews, *Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy*, 5 STAT. SURV. 1 (2011).

246. See Dwork, *supra* note 27.

individual level include privacy policy “nutrition labels”<sup>247</sup> and personal data stores.<sup>248</sup>

The number, variety, and domain of application of these privacy principles, guidelines, methods and systems are expansive. This poses a substantial challenge for policymakers, scholars, and practitioners alike because there is little formal guidance for selecting among these privacy enhancing methods and systems, or for evaluating the privacy considerations related to a particular data release case. This situation contrasts starkly with the related field of information security, which boasts well-known, regularly updated catalogs of threats, vulnerabilities, and controls organized within well-defined categories. By comparison, information privacy literature describes many controls, threats, vulnerabilities, and measures of utility, but no catalog or ready categorizations exist for assessing privacy-related factors.<sup>249</sup>

#### A. CHARACTERIZING PRIVACY CONTROLS, THREATS, VULNERABILITIES, AND USES

We propose a framework, modeled on the use of categorizations and catalogs in information security, that can be used to evaluate specific cases of government data releases, identify privacy concerns, and develop privacy-improving approaches that are appropriate for a specific case. This framework distinguishes between privacy controls, threats, harms, vulnerabilities, and utility:

- Privacy *controls* (interventions) are defined as methods or mechanisms that can be applied within a particular data release case to enhance privacy and confidentiality.

---

247. Patrick Gage Kelley et al., *A “Nutrition Label” for Privacy*, 5 SYMP. ON USABLE PRIVACY & SECURITY, Article No. 4 (2009).

248. See, e.g., Yves-Alexandre de Montjoye et al., *openPDS: Protecting the Privacy of Metadata through SafeAnswers*, PLOS ONE (July 9, 2014), <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098790>.

249. This Article distinguishes between security and privacy controls in line with how these terms are used in NIST guidelines. Security controls encompass safeguards within information systems and their environments to protect information during processing, storage, and transmission. Categories of security controls include access, awareness and training, audit and accountability, identification and authentication, maintenance, risk assessment, and system and information integrity controls. Privacy controls are administrative, technical, and physical safeguards to protect and ensure the proper handling of information associated with privacy risks. Categories of privacy controls include authority and purpose, accountability and audit, risk management, data quality and integrity, data minimization and retention, individual participation and redress, security, transparency, and use limitation controls. See NIST, CONTROLS (final), *supra* note 228.

The term control is inclusive, encompassing more generally targeted interventions, such as privacy education, as well as information security controls like encryption, traditional procedural controls such as certification of authorized users, statistical disclosure limitation methods such as data perturbation,<sup>250</sup> and legal controls such as criminal penalties.

- Privacy *threats* are defined broadly as potential adverse circumstances or events that could cause harm to a data subject as a result of the inclusion of that subject's data in a specific data collection, storage, management, or release.<sup>251</sup> Threats are broadly inclusive, and meant to encompass everything from government surveillance, to accidentally leaving backup tapes on a bus, to natural disasters.
- Privacy *harms* are defined as injuries, such as embarrassment, reputational loss, loss of employability or insurability, imprisonment, or death, sustained by data subjects as a result of the realization of a threat.<sup>252</sup>
- Privacy *vulnerabilities* are defined as characteristics that increase the likelihood that threats will be realized.<sup>253</sup> These characteristics are defined as broadly inclusive, encompassing characteristics of the data; of the systems used to collect, store, manage or release the data; and of

---

250. Data perturbation refers to the masking of data using techniques such as random noise addition, random or controlled rounding of values, or swapping of values. For an overview of such techniques, see FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16.

251. Note that this is compatible with, but more broadly defined than the concept of a *threat model*. A threat model, depending on the field in which it is characterized, typically involves identification of the category of cause (e.g., natural disaster, human error, malicious behavior) potentially leading to the bad outcome, and characterization of the extent of that cause (e.g., the background knowledge and capability of an attacker).

252. For a discussion of the broad range of privacy harms, see Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477 (2006).

253. Note that this definition is analogous to the definition of vulnerability within information security, but distinct in that information security vulnerabilities identify specific system flaws in providing a defined property of information assurance. The motivation for the more general definition of privacy vulnerability is that formal definitions of privacy assurance properties are neither complete nor comprehensively accepted, and thus the notion of complete assurance, and the complementary notion of a flaw or defect, are not well defined.

the related context in which these systems operate and in which interactions with these systems occur.

- *Utility* is defined broadly as the analytic value of the data. It describes the types of analyses that the data can support. The use of certain privacy controls, such as traditional statistical disclosure limitation techniques, can greatly diminish the utility of the data in practice.<sup>254</sup>

We believe this Article is the first to adopt this categorization explicitly and to use the specific definitions above.<sup>255</sup> However, elements of this categorization are closely related not only to information security definitions, as mentioned, but also to a line of work in several other fields.<sup>256</sup>

To aid such an analysis, our proposed framework divides data releases into multiple stages based on a lifecycle model of government data release. A fully developed lifecycle model, as used frequently in information

254. Note that utility is not an explicit part of standard information security frameworks. Instead, information security effectively defines utility as the maintenance of security properties of the system, such as integrity, secrecy, availability, and non-repudiation.

255. This may be considered a formalization of the framework we and our collaborators sketch in prior work. *See, e.g.*, Alexandra Wood et al., Comments to the Department of Health and Human Services, Federal Policy for the Protection of Human Subjects; Proposed Rules, Docket No. HHS-OPHS-2015-0008 (Jan. 6, 2016), [http://privacytools.seas.harvard.edu/files/privacy\\_tools\\_project\\_response\\_to\\_common\\_rule\\_nprm.pdf](http://privacytools.seas.harvard.edu/files/privacy_tools_project_response_to_common_rule_nprm.pdf); Salil Vadhan et al., *supra* note 24; Micah Altman et al., Comments to the White House Office of Science and Technology Policy Re: Big Data Study; Request for Information (Mar. 31, 2014), <http://privacytools.seas.harvard.edu/files/privacytools/files/whitehousebigdataresponse1.pdf>.

256. *See, e.g.*, WILLENBORG & DE WAAL, *supra* note 136 (explicitly characterizing the primary privacy threat models for releases of official statistics); Cynthia Dwork, *Differential Privacy*, 33 INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES & PROGRAMMING, PT. II, at 1 (2006) (defining differential privacy in terms of a specific combination of threat model, vulnerability characterization, and choice of control); Wu, *supra* note 23 (providing an overview and detailed analysis of threat models used in the privacy context); NIST, *supra* note 227 (proposing draft guidance to characterize privacy in terms of controls, threats, and risks, using somewhat narrower definitions than those we adopt); Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. (forthcoming 2016) (N.Y.U. Pub. L. & Legal Theory Working Paper No. 530), [http://lsr.nellco.org/nyu\\_plltwp/530](http://lsr.nellco.org/nyu_plltwp/530) (arguing that “data release policy should look more like data security policy”); Adam D. Thierer, *A Framework for Benefit-Cost Analysis in Digital Privacy Debates*, 20 GEORGE MASON L. REV. 1055 (2013) (describing a high-level abstract cost-benefit analysis that includes references to the concept of risk, vulnerabilities, and controls, although these concepts are neither explicitly defined, nor a central part of the analysis).

science and in records management,<sup>257</sup> documents the information objects, actors, action space, and incentives across each stage of information collection, processing, and use. Moreover, frameworks such as privacy by design, and laws such as CIPSEA, as discussed above, advocate using lifecycle analysis for data management generally, although they do not provide specific guidelines for doing so.

In contrast to existing approaches to lifecycle management of privacy risks, we apply the information lifecycle not as a design principle but as a way of decomposing the privacy risks, actors, and potential interventions. Further, we have adapted the stages of the research information lifecycle (Figure 1) to match the phases of activity and areas of regulatory concern that are associated with the government data release cases discussed in Part II.

In the remainder of Part III, we develop a framework for this catalog, sketch its contours, and populate selected portions of its contents. We start by developing a categorization system for privacy controls and then show how this categorization scheme can be applied and expanded to characterize intended uses, privacy threats, and privacy vulnerabilities. In Section III.D and Part IV, we offer some suggestions for selecting controls for a particular data release case based on the uses, threats, and vulnerabilities of the release.

---

257. Our proposed framework incorporates a partial “lifecycle” model that focuses on the stages of activity associated with government data releases. Lifecycle models have been used in biology for at least a hundred years. They have been applied to processes in many fields, such as project management and software development, and as a general idea, lifecycle models have been previously applied to privacy analysis. Notably, one of the principles of privacy by design is to provide full lifecycle security. Formal models of the lifecycle of information are a more recent development, however, and we base our lifecycle model (Figure 1) on existing models developed for the curation of research information. *See, e.g.*, Micah Altman, *Mitigating Threats to Data Quality Throughout the Curation Lifecycle* (position paper from a workshop, Curating For Quality: Ensuring Data Quality to Enable New Science, Arlington, Virginia, Sept. 10–11, 2012), <http://datacuration.web.unc.edu>; Sarah Higgins, *The DCC Curation Lifecycle Model*, 3 INT’L J. DIGITAL CURATION 134 (2008), <http://www.dcc.ac.uk/resources/curation-lifecycle-model>. A somewhat novel feature of information lifecycles is that the object of concern, information, can be viewed as both a conceptual (e.g., measurements describing a subject) and a logical entity (e.g., a particular computer file containing those measurements). Further, the latter is easily replicated, and one copy of the same file may be retained while another is accessed or distributed. Thus models of information lifecycle differ from information flow, where the latter is concerned primarily with the storage and transmission of information and the grouping the types of actors and actions to which the conceptual information entities are subject.

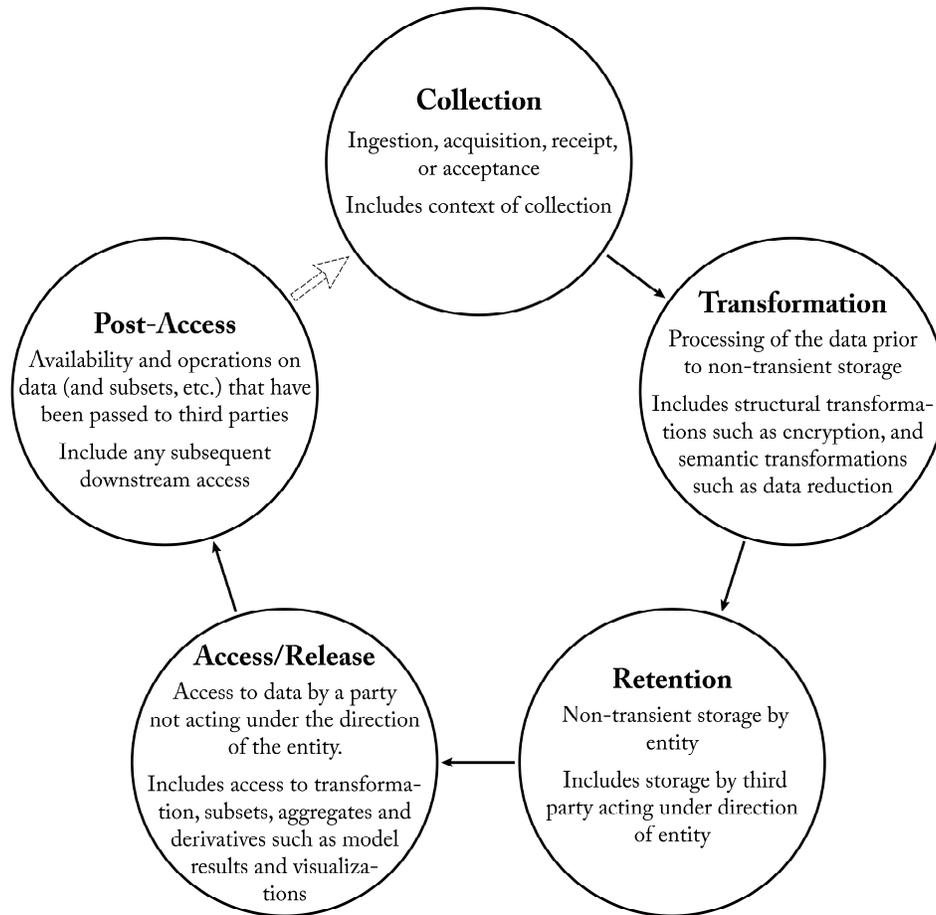


Figure 1: Lifecycle model for government data releases, based on use cases in Part II.

## B. DEVELOPING A CATALOG OF PRIVACY CONTROLS AND INTERVENTIONS

Policy researchers, scholars, and privacy advocates have suggested scores of controls and interventions to improve privacy protection, ranging from the voluntary use of icons to communicate privacy policies, to giving data subjects rights to sue, to storing data in subject-controlled vaults, to performing all analyses only upon data encrypted at collection. In addition, information security catalogs list dozens more controls that are aimed at enhancing the protection of data managed within information systems. A policymaker or manager of a data release program is tasked with determining how to approach such complexity when designing a data release that protects the privacy interests of the subjects of the data.

A classification of controls is clearly needed to provide guidance. Of information security standards, FISMA<sup>258</sup> and the implementing guidelines<sup>259</sup> from NIST provide the most systematic and extensive classifications of controls. Moreover, with NIST's latest draft guidelines,<sup>260</sup> these standards would become one of the few to address privacy controls explicitly. FISMA's catalog of controls includes the following: accountability, audit, and risk management controls such as impact and risk assessments; data quality management and integrity monitoring; data minimization and retention; individual participation and redress; transparency; and use limitations.<sup>261</sup>

These privacy controls provide a useful start, and they should be considered when designing a data release policy. However, this list is far from comprehensive. For example, it excludes many of the more modern statistical and computational approaches to protecting privacy. Moreover, FISMA has important scope limitations. It focuses on controls implemented through technical and procedural means and those that are implemented within an existing agency policy, not on controls that could be selected when designing a policy for data release.

Since the design of policies for data release is our main concern, our catalog expands the scope of controls to consider controls implementable through the entire range of means available to policy makers. We categorize the space of privacy controls as follows:

- *Procedural means*, defined broadly as adopting procedures internal to an organization, such as implementing notice, creating inventories, or vetting internal and external access to databases;
- *Technical means*, defined broadly to include statistical methods, computational methods such as encryption, and human factors analysis such as readability analysis of privacy policies;
- *Educational means*, defined broadly to include any intervention intended to inform data subjects, data controllers, and data recipients that interact with the

---

258. Federal Information Security Management Act, 44 U.S.C. §§ 3541–3549 (2012).

259. See, e.g., NIST, STANDARDS, *supra* note 228; NIST, FIPS PUB. 200, MINIMUM SECURITY REQUIREMENTS FOR FEDERAL INFORMATION AND INFORMATION SYSTEMS (2006); NIST, SPECIAL PUB. 800-18, GUIDE FOR DEVELOPING SECURITY PLANS FOR FEDERAL INFORMATION SYSTEMS (2006).

260. NIST, *supra* note 227.

261. See NIST, CONTROLS (final), *supra* note 228.

system; data subjects, controllers, or recipients generally; or the public at large about privacy practices and risks;

- *Economic means*, defined broadly as including any intervention intended to change the economic incentives of the stakeholders, such as the imposition of fees or fines, or the provision of insurance; and
- *Legal means*, defined specifically as interventions intended to change the legal rights of or relationships among stakeholders, such as safe harbor provisions, or private rights of action.

Policymakers should consider the appropriate staging of policy interventions and the means at their disposal for constructing these interventions. The review below discusses many of the most commonly applied controls, and some promising new approaches from the literature, for releasing government data about individuals in a privacy-protective way. It is not intended to be exhaustive; rather, it is illustrative of the spectrum of procedural, economic, educational, legal, and technical approaches available, and how they interact with one another, at each stage of the information lifecycle.

#### 1. *Privacy Controls at the Collection and Acceptance Stage*

The first stage of the lifecycle for government data releases begins with collection of the data. This Article uses the term collection broadly to include acceptance, ingestion, acquisition, or receipt of data. Controls applied at this stage typically affect what is collected, the manner in which it is collected, and the context of collection. This Article reviews some common controls, and some that demonstrate the range of approaches available.

Notice and consent are cornerstones of the fair information practice principles. They have been, and will continue to be, a common tool for protecting privacy. To improve notice, commentators have proposed public education initiatives to inform citizens of the types of data collected, how they are used, and the privacy risks associated with government data programs. Such initiatives may include practical demonstrations of government data uses or of the types of reidentification attacks that could be employed.<sup>262</sup> Consent mechanisms are evolving, and there is movement in some areas towards more portable and broader

---

262. See, e.g., Jeff Jonas & Jim Harper, *Open Government: The Privacy Imperative*, in *OPEN GOVERNMENT* (Daniel Lathrop & Laurel Ruma eds., 2010).

consent for certain uses of information, such as research uses.<sup>263</sup> In particular, consent to data collection may not be a sufficient mechanism for privacy protection. Privacy policies are widely considered to be too complex for individuals to readily understand, and, in some cases, the summaries of the policies provided by data collectors are inaccurate.<sup>264</sup> Standard policies often do not clearly convey the permitted third party uses and disclosures of personal information, allow individuals to consent to only certain uses or uses by specific parties, or enable individuals to modify or revoke their consent over time.<sup>265</sup> There is a growing recognition that consent should not be treated simply as a binary action that occurs at the time of data collection and functions to restrict collection, but as a continual process in which the subject agrees to collection, retention, transformation, access, and post-access uses and controls, within a defined context. To address these and related issues around consent, scholars have proposed alternative tools to standardize privacy policies and simplify their terms using, for example, icons or “nutrition labels.”<sup>266</sup> At the same time, requiring consent from individuals may reduce participation in a data collection program and thereby reduce the quality of the data collected,<sup>267</sup> though participation can be incentivized by offering payments to individuals who agree to share their information.<sup>268</sup> The costs of operating more effective consent programs that allow for more granular permissions, or that provide payments to data subjects, can be shared with the data user by charging fees to access the data.<sup>269</sup>

---

263. See Effy Vayena et al., *Caught in the Web: Informed Consent for Online Health Research*, SCI. TRANSLATIONAL MED., Feb. 20, 2013, at 1.

264. See Lorrie Faith Cranor, *Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice*, 10 J. ON TELECOMM. & HIGH TECH. L. 273 (2012).

265. See KIERON O'HARA, TRANSPARENT GOVERNMENT, NOT TRANSPARENT CITIZENS: A REPORT ON PRIVACY AND TRANSPARENCY FOR THE CABINET OFFICE 53 (2011).

266. See, e.g., Renato Iannella & Adam Finden, *Privacy Awareness: Icons and Expression for Social Networks*, Proceedings of the 8th Int'l Workshop for Virtual Goods (2010); Gage Kelley et al., *supra* note 247; AZA RASKIN & ARUN RANGANATHAN, PRIVACY: A PICTOGRAPHIC APPROACH, in W3C WORKSHOP ON PRIVACY FOR ADVANCED WEB APIS (2010), <http://www.w3.org/2010/api-privacy-ws/papers/privacy-ws-22.txt>; *The Platform for Privacy Preferences 1.0 (P3P1.0) Specification*, W3C (Apr. 16, 2002), <http://www.w3.org/TR/2002/REC-P3P-20020416>.

267. See O'HARA, *supra* note 265, at 49–50.

268. See Bart van der Sloot, *On the Fabrication of Sausages, or of Open Data and Private Data*, E JOURNAL EDEMOCRACY & OPEN GOV'T 136 (2011).

269. See O'HARA, *supra* note 265, at 49–50.

In addition to notice and consent, agencies often seek to provide privacy protection at the acceptance stage by implementing measures to operationalize several other fair information practice principles: collection limitation, data minimization, and purpose specification in the design of a data collection program.<sup>270</sup> For instance, governments may prohibit the collection of personal information except for specific, limited purposes.<sup>271</sup> In these settings, governments may require an agency to specify and document the purpose of collection, which can be referenced when auditing for data misuses.<sup>272</sup> Organizations may also appoint a data protection officer or chief privacy officer who oversees the collection, storage, use, and dissemination of personal data to ensure that practices are consistent with the fair information practice principles.

Another common mechanism for privacy protection in data collection is oversight by a privacy board, institutional review board, or other independent panel. For example, researchers who receive federal funding to conduct a study involving human subjects must secure approval from an institutional review board and follow procedures for informing the subjects of the benefits and risks, including privacy risks, related to their participation in the study; specifying the nature, scope, and purpose of the study; and obtaining subjects' consent to participation.<sup>273</sup> The scope of the research and future uses of the data is limited to the activities described in the consent form. Some studies use consent procedures that enable subjects to grant permission for certain uses but not others, and involve frequent follow-up meetings during which new consent forms can be signed to authorize research in additional areas. In other cases, it can be cost prohibitive or otherwise unfeasible to contact all of the participants in a research study and obtain consent for new uses of their personal information. Violations of any of these protocols can lead to the withdrawal of federal research funding if backed by regulatory enforcement mechanisms.

Privacy impact assessments are frequently cited as a recommended tool for balancing utility and privacy and for choosing appropriate privacy safeguards when collecting, storing, using, and disseminating personal information.<sup>274</sup> All federal executive agencies are required to conduct

---

270. See U.S. DEPT OF HEALTH, EDUC., & WELFARE, *supra* note 228.

271. See Scassa, *supra* note 19.

272. See O'HARA, *supra* note 265, at 29.

273. See 45 C.F.R. pt. 46 (2014).

274. See, e.g., Francesco Molinari & Jesse Marsh, *Does Privacy Have to Do with Open Data? Some Preliminary Reflections—And Answers*, in PROCEEDINGS OF THE CEDEM13 CONFERENCE (Peter Parycek & Noella Edelmans eds., 2010); Ugo Pagallo & Eleonora

privacy impact assessments for information technology systems containing personally identifiable information.<sup>275</sup> Such assessments vary between agencies but typically involve a review of the nature and source of the information to be collected, the purpose and intended use of the information to be collected, the intended recipients of the information, the rights of individuals to consent to or decline to provide their information, and the security controls to be used.<sup>276</sup> Note, however, that such assessments do not generally involve documenting specific privacy threats or vulnerabilities. Section III.C details this shortcoming.

## 2. *Privacy Controls at the Transformation Stage*

Transformation of data encompasses a range of alterations. Transformations may be structural or semantic, and transformations may be lossy or lossless. Transformation may be applied at multiple stages, including directly after collection and prior to long term retention, after a substantial retention period and prior to access, or integrated with access. Applying transformations earlier provides greater protection, but may limit the range of analysis that may be performed later. For example, the common transformation of redacting or aggregating information can be employed any time after collection until release. If applied immediately after collection, redacting or aggregating information reduces the harm expected in the case of a data breach; however, doing so also curtails the potential to link, merge, or update the data.

Transformations applied in early stages typically involve public- or private-key encryption.<sup>277</sup> Standard forms of private- and public-key encryption mitigate disclosure risks from breaches during data retention. Encryption approaches to transformation are typically non-lossy; the original information can be obtained in its entirety given access to a complete set of encryption keys, which may be divided across stakeholders.<sup>278</sup> Other approaches to transformation typically cause information loss. The most common approach to sanitization or de-identification is to manually review the fields in a set of data and remove

---

Bassi, *Open Data Protection: Challenges, Perspectives, and Tools for the Reuse of PSI*, in DIGITAL ENLIGHTENMENT Y.B. 2013 (M. Hildebrandt et al. eds., 2013).

275. See OFFICE OF MGMT. & BUDGET, *supra* note 217.

276. See *id.*

277. For a detailed description of public- and private-key encryption standards for federal government information systems, see NIST, FIPS PUB. 140-2, SECURITY REQUIREMENTS FOR CRYPTOGRAPHIC MODULES (2001), <http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf>.

278. See Hugo Krawczyk, *Secret Sharing Made Short*, 13 ANN. INT'L CRYPTOLOGY CONF. 136 (1994).

direct and indirect identifiers.<sup>279</sup> Fields are typically redacted, according to varying standards such as the HIPAA Privacy Rule safe harbor de-identification standard,<sup>280</sup> based on the type of information, the intended recipients, the potential uses of the data, the regulatory requirements, and best practices in the relevant industry. Transformation methods derived from traditional statistical disclosure limitation are typically applied at post-retention stages and include aggregation, suppression, and perturbation.<sup>281</sup> However, simple methods such as removing personally identifiable information or masking data through aggregation and perturbation of individual points are generally insufficient when it comes to large datasets, short of rendering the data useless.<sup>282</sup>

Another common privacy control is aggregation or the production of summary statistics, such as contingency tables or tables that provide the frequencies of co-occurring attributes. For example, a three-dimensional contingency table based on census data for Norfolk County, Massachusetts, might have an entry listing how many people in the population are female, under the age of forty, and rent their home. Data may also be released using data visualizations, which are graphical depictions of a dataset's features or statistical properties. Data visualizations are especially useful for comprehending large amounts of data, perceiving emergent properties, identifying anomalies, understanding features at different scales, and generating hypotheses.<sup>283</sup>

Another approach is to generate synthetic data from a statistical model that has been developed using the original dataset. Methods for generating synthetic data were first developed for filling in missing entries, and are now considered attractive for protecting privacy because a synthetic dataset does not directly refer to any "real" person.<sup>284</sup> They are, however, of

---

279. See Thomas P. Keenan, *Are They Making Our Privates Public? Emerging Risks of Governmental Open Data Initiatives*, in *PRIVACY & IDENTITY MANAGEMENT FOR LIFE* 1, 12 (Jan Camenisch et al. eds., 2012).

280. 45 C.F.R. § 164.514(b) (2014).

281. See FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 16.

282. See Alan F. Karr & Jerome P. Reiter, *Using Statistics to Protect Privacy*, in *PRIVACY, BIG DATA, AND THE PUBLIC GOOD* 276 (Julia Lane et al. eds., 2014).

283. See COLIN WARE, *INFORMATION VISUALIZATION: PERCEPTION FOR DESIGN* (3d ed. 2013); Frank D. McSherry, *Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis*, Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (2009).

284. See John M. Abowd & Lars Vilhuber, *How Protective Are Synthetic Data?*, in *PRIVACY IN STATISTICAL DATABASES* (Josep Domingo-Ferrer & Yucel Saygin, eds., 2008); Stephen E. Fienberg, *Conflicts Between the Needs for Access to Statistical Information and Demands for Confidentiality*, 10 J. OFFICIAL STAT. 115 (1994); Donald B. Rubin, *Discussion of Statistical Disclosure Limitation*, 9 J. OFFICIAL STAT. 461 (1993).

limited use because only the properties that have been specifically modeled are present in the synthetic dataset. For example, a synthetic dataset designed to accurately reproduce the univariate means and correlations of the original data may not yield the same results when non-linear models are estimated.

The transformation method choice should be made after careful consideration of the privacy guarantee that is required. In some cases involving information deemed to be benign, it may not be necessary to use a transformation that satisfies a strong guarantee of privacy. In other cases where privacy concerns are high, it may be necessary to use an advanced aggregation, perturbation, or synthetic data algorithm that satisfies a formal notion of privacy known as differential privacy,<sup>285</sup> to produce a dataset that can be shared widely. The transformation decision should also take into account the analyses that must be supported by the data release, as the techniques employed for reducing disclosure risks can affect potential uses and analyses.<sup>286</sup> In addition, such controls should be implemented in consultation with experts, as improper design can substantially reduce the privacy and utility of a data release. For example, when New York City officials de-identified taxi trip data prior to release in 2014, they used an ineffective technique (a simple hash function) that made discovery of the hack license and medallion numbers of all of the taxi drivers quite easy.<sup>287</sup> In another case, researchers discovered errors in de-identified public use datasets published by the U.S. Census Bureau between 2000 and 2007, with analytical results varying by as much as 15% from the actual statistics due to misapplication of statistical disclosure limitation techniques.<sup>288</sup> Regardless of the transformation technique chosen, an organization should be transparent about its transformation practices, for instance by providing details in the metadata associated with the data, so that users of the data will be informed about potential limitations of the data.<sup>289</sup>

---

285. See Dwork, *supra* note 27.

286. See, e.g., Kingsley Purdam & Mary Elliot, *A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records*, 39 ENV'T & PLAN. A 1101 (2007).

287. See Goodin, *supra* note 87.

288. See J. Trent Alexander et al., *Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications*, 74 PUB. OPINION Q. 551 (2010).

289. See O'HARA, *supra* note 265, at 77.

### 3. *Privacy Controls at the Retention Stage*

We define retention broadly to include any form of non-transient storage by the data controller or a party acting under the controller's direction. Information security controls already focus heavily on the retention phase, and so this Article summarizes controls here without providing a detailed discussion. A number of information security controls are common at the retention stage, such as access control, maintenance, security assessments, authentication procedures, incident monitoring and response, and audits.<sup>290</sup> For example, for some categories of confidential data, industry standards may require encryption,<sup>291</sup> or laws may require encryption where reasonable.<sup>292</sup> Organizations commonly implement data retention and decommissioning policies to ensure data are retained for no longer than necessary and data backups are destroyed after a certain length of time.<sup>293</sup> Many states require personal information maintained by state agencies or businesses to be destroyed when the data are no longer needed.<sup>294</sup> In addition, data sharing agreements often specify that the recipient must destroy the data within some period, such as one year after receipt, and law may also require such a contractual provision.<sup>295</sup>

Data policies may also include data integrity and accuracy provisions. For example, data policies may require organizations to keep data accurate and up to date, ensure that individuals can access and correct data about themselves, and notify third party data recipients of any discovered inaccuracies in delivered data.<sup>296</sup> Data tethering can operationalize such policies. Data tethering ensures that all instances of a piece of information are linked, so that changes in one place are reflected in all copies of the data.<sup>297</sup>

Privacy dashboards and personal data stores are tools which individuals can use to express detailed permissions regarding retention and use of their data. An individual can use a web-based privacy dashboard to grant

290. *See, e.g.*, NIST, CONTROLS (final), *supra* note 228.

291. *See, e.g.*, SECURITY STANDARDS COUNCIL, PAYMENT CARD INDUSTRY DATA SECURITY STANDARD: REQUIREMENTS AND SECURITY ASSESSMENT PROCEDURES (2015).

292. *See, e.g.*, Health Insurance Portability and Accountability Act Security Rule, 45 C.F.R. §§ 164.312(a)(2)(iv), (e)(2)(ii) (2014).

293. *See* Jonas & Harper, *supra* note 262, at 324.

294. *See, e.g.*, N.J. Stat. § 56:8-162 (2006); Mo. Stat. § 288.360.

295. *See, e.g.*, FERPA, 34 C.F.R. § 99.31(a)(6)(iii)(C) (2014).

296. *See, e.g.*, Privacy Act of 1974, 5 U.S.C. § 552a(d)(1) (2012); van der Sloot, *supra* note 268, at 144.

297. *See* Scassa, *supra* note 19.

granular access permissions to her data only to select parties or for select uses.<sup>298</sup> Personal data stores enable individuals to effectively exercise fine-grained control over where information about them is stored and how it is accessed, and thus choose to share specific pieces of personal information at specific times with specific parties.<sup>299</sup> Personal data stores not only provide increased control but, as user-controlled, interactive systems, are a potential foundation for developing richer accountability mechanisms, online aggregation methods, and advanced security mechanisms.

Transparency, legal, and technical controls may also be available at the retention stage. An example of a transparency intervention at this stage is a data asset register, which discloses to the public what data are maintained by an organization.<sup>300</sup> Legal interventions include statutory breach reporting requirements, which require organizations to notify individuals and enforcement bodies in the event of a data security breach.<sup>301</sup> Examples of technical measures include federated databases, for enabling controlled queries across databases maintained by different organizations,<sup>302</sup> computable policies, for automating the enforcement of privacy policies,<sup>303</sup> and secret sharing and other techniques for managing keys for encrypted systems.<sup>304</sup>

#### 4. *Privacy Controls at the Release and Access Stage*

Many controls are applied at the release stage. We define release inclusively to mean access to any transformation, subset, or derivative of the data by a party not acting under the direction of the data controller. Broadly, controls applied at the access stage may affect what portions of data are accessed, how decisions to grant access are made, and the

---

298. See van der Sloot, *supra* note 268, at 149.

299. See Tom Kirkham et al., *The Personal Data Store Approach to Personal Data Security*, 11 IEEE SECURITY & PRIVACY 12 (2013); see, e.g., Yves-Alexandre de Montjoye et al., *On the Trusted Use of Large-Scale Personal Data*, 35 IEEE DATA ENG. BULL. 5 (2013).

300. See, e.g., OFFICE OF MGMT. & BUDGET, *supra* note 195.

301. See generally NATIONAL CONFERENCE OF STATE LEGISLATURES, *Security Breach Notification Laws*, <http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx> (last visited July 15, 2015) (listing breach notification laws by state).

302. See DATA PRIVACY AND INTEGRITY ADVISORY COMMITTEE, *PRIVACY POLICY AND TECHNOLOGY RECOMMENDATIONS FOR A FEDERATED INFORMATION-SHARING SYSTEM*, Report No. 2011-01 (2011), [http://www.dhs.gov/xlibrary/assets/privacy/privacy\\_dpiac\\_report\\_2011\\_01.pdf](http://www.dhs.gov/xlibrary/assets/privacy/privacy_dpiac_report_2011_01.pdf).

303. See, e.g., Lalana Kagal & Joe Pato, *Preserving Privacy Based on Semantic Policy Tools*, 8 IEEE Security & Privacy 25 (2010).

304. See, e.g., Adi Shamir, *How to Share a Secret*, 22 COMM. ACM 612 (1979).

conditions imposed upon those accessing the data. In some cases additional transformation, such as data aggregation, is integrated into the access phase.

Operational policy, a central component of any data management program, can embed privacy controls at the release stage. When releasing information, a government agency must make a decision regarding the proper balancing of privacy and transparency. For example, courts have historically made their records available to the public under a very strong presumption of openness,<sup>305</sup> while statistical agencies have required strong confidentiality protections for their data and taken a more conservative approach.<sup>306</sup> Governments are increasingly pressured to make information available under a presumption of openness,<sup>307</sup> and commentators have suggested that expert panels, including a broad range of stakeholders, be involved in developing policies for making release decisions.<sup>308</sup> Risk assessments and checklists are also used to guide an evaluation of the privacy risks associated with a set of data, to help balance privacy and utility considerations, and to determine an appropriate release mechanism or privacy control to mitigate these risks.<sup>309</sup>

Organizations also use access controls when sharing data through an information system. Such a system may require all users to register and provide contact information before accessing the data, and it may also employ authentication protocols to verify the identity of an individual. Organizations can also use tiered access systems to grant different levels of access to different parties based on, for example, the affiliations or credentials of the individual. Tiered access may also incorporate more advanced data sharing models. For instance, aggregate statistics in the form of a contingency table might be provided to the public. An interactive query system might be made available to a community of researchers, and raw data might be made available to a small number of analysts who are approved through a careful screening process.

---

305. See Conley et al., *supra* note 19, at 778.

306. See, e.g., Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), 44 U.S.C. § 3501 note (2012).

307. See, e.g., Exec. Order No. 13,642, 3 C.F.R. 244 (2014) (Making Open and Machine Readable the New Default for Government Information), <https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government->; O'HARA, *supra* note 265, at 73.

308. See, e.g., Katleen Janssen & Sara Hugelier, *Open Data: A New Battle in an Old War Between Access and Privacy*, in DIGITAL ENLIGHTENMENT Y.B. 2013 (M. Hildebrandt et al., eds., 2013); O'HARA, *supra* note 265.

309. See Pagallo & Bassi, *supra* note 274.

Large data repositories and statistical agencies like the U.S. Census Bureau use secure data enclaves to control access to and use of sensitive information. A physical or virtual data enclave is a secure environment that enables authorized users to access confidential data and analyze the data using provided statistical software such as R, Stata, or SAS. A researcher must apply for access, typically by providing proof of identity, describing the scope and methodology of the proposed research, establishing the need for non-public data and the benefit of conducting the research, demonstrating research expertise or specialized knowledge, and, if applicable, agreeing to be bound by the federal confidentiality laws and penalties that apply to agency employees.<sup>310</sup> The secure data enclave controls and tracks all activity by the researcher, limits the linkages that can be made to auxiliary data, and maintains records that can later be audited by a third party. The data cannot be removed from the secure environment, and any generated tables, model coefficients, or other results are vetted for disclosure risks prior to publication.<sup>311</sup> Secure enclaves hosted by federal statistical agencies have not led to any known security breaches, but their use makes it difficult to validate and replicate research results.

Interactive mechanisms are systems that enable users to submit queries about a collection of data and receive only the results of the query analysis, perhaps rendered in the form of a table or visualization. A dataset is stored securely and a user is never given direct access to the raw data. Rather, a curator mediates access. Such systems can restrict access to queries that are associated with greater privacy risks, and they potentially allow for very sophisticated queries. For example, the Census Bureau's online Advanced Query System allows users to create their own customized contingency tables.<sup>312</sup>

Many of these privacy controls, including privacy-aware methods for contingency tables, synthetic data, data visualizations, and interactive mechanisms, have been successfully used to share data while protecting privacy, with no serious compromises discovered to date.<sup>313</sup> The fact that these systems do not provide direct access to raw data does not automatically ensure privacy, but when made privacy-aware in an appropriate way, they can provide strong protection. Further, many of

---

310. *See, e.g.*, Penn State Research Data Center, Applying for Special Sworn Status, <http://www.psurdc.psu.edu/content/applying-special-sworn-status> (last visited May 28, 2015).

311. *See, e.g.*, U.S. CENSUS BUREAU, *supra* note 153.

312. U.S. CENSUS BUREAU, *supra* note 151.

313. *See* Salil Vadhan et al., *supra* note 24.

these forms of data sharing have even been shown to be compatible with a strong new privacy guarantee known as differential privacy.<sup>314</sup> Differential privacy provides a framework for measuring and reducing the risk of disclosing privacy-sensitive information about individuals when analyzing and sharing data.<sup>315</sup> An appropriately designed differentially private system can provide strong, provable guarantees that individual-specific information will not leak, regardless of what auxiliary information may be available, while still allowing for rich statistical analysis of a dataset.<sup>316</sup>

Secure multiparty computations are electronic protocols that enable two or more parties to carry out a computation that involves both of their datasets in such a way that no party needs to explicitly hand a dataset to any of the others.<sup>317</sup> Because secure multiparty computation allows for queries to be computed without the need for all data storage to be centralized, it reduces the harm from data breach, and allows computations across parties that do not fully trust each other.<sup>318</sup> In theory, it can be combined with the interactive mechanisms and privacy aware computational methods previously mentioned.<sup>319</sup>

Other advanced encryption approaches can enable computations on data while limiting learning about the underlying data. Techniques from cryptography can ensure that no party learns anything beyond the result of the computation. For example, functional or homomorphic encryption is

---

314. For the foundations of differential privacy, see Irit Dinur & Kobbi Nissim, *Revealing Information while Preserving Privacy*, in PROCEEDINGS OF THE 22ND ACM SIGMOD-SIGACT-SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS 202 (2003); Cynthia Dwork & Kobbi Nissim, *Privacy-Preserving Datamining on Vertically Partitioned Databases*, in PROCEEDINGS OF THE 24TH ANNUAL INTERNATIONAL CRYPTOLOGY CONFERENCE 528 (2004); Avrim Blum, Cynthia Dwork, Frank McSherry, & Kobbi Nissim, *Practical Privacy: The SuLQ Framework*, in PROCEEDINGS OF THE 24TH ACM SIGMOD-SIGACT-SIGART SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS 128 (2005); Cynthia Dwork, Frank McSherry, Kobbi Nissim, & Adam Smith, *Calibrating Noise to Sensitivity in Private Data Analysis*, in PROCEEDINGS OF THE 3RD CONFERENCE ON THEORY OF CRYPTOGRAPHY 265 (2006).

315. See Dwork, *supra* note 256.

316. See Dwork, *supra* note 27.

317. See Yehuda Lindell & Benny Pinkas, *Secure Multiparty Computation for Privacy-preserving Data Mining*, 1 J. PRIVACY & CONFIDENTIALITY 59 (2009); Alan F. Karr et al., *Secure Regression on Distributed Databases*, 14 J. COMPUTATIONAL & GRAPHICAL STAT. 263 (2005).

318. See sources cited *supra* note 317.

319. See, e.g., Amos Beimel, Kobbi Nissim & Eran Omri, *Distributed Private Data Analysis: Simultaneously Solving How and What*, 28 ANN. INT'L CRYPTOLOGY CONF. 451 (2008) (exploring the combination of secure multiparty computation and differential privacy).

an encryption method being developed to enable computations to be performed on encrypted data without decrypting the data first and exposing it to attack.<sup>320</sup>

Interventions at this stage may also have a transparency or economic component. For example, data asset registers, transparency panels, and open debates (with published minutes) can inform the public about what types of information governments hold and release and how they decide which data to release to or withhold from the public.<sup>321</sup> Charging fees, either one-time or subscription, or otherwise raising the costs of access may discourage individuals from accessing the data for improper purposes. For example, a recent proposal for tiered access to court records would make sanitized versions of court records available online and unsanitized court records available only on-site at the courthouse, as a way to limit the aggregation and circulation of sensitive information while maintaining utility to members of the public.<sup>322</sup>

#### 5. *Privacy Controls at the Post-Access Stage*

Once data are released or exit from a formal information system, the set of controls that can be effectively applied changes, and the privacy risks continue to evolve. In some ways, the post-access phase resembles the pre-collection phase, as private information is now available outside the data controller's system and available for use or even re-collection. However, there are substantial distinctions in that the information at the post-access stage may have been transformed, that different rights and responsibilities may be attached to it, and that the data subject is much less likely to be involved in decisions over re-collection or reuse (in the absence of specific interventions to ensure this involvement).

Privacy risks arising from data release change over time. Subsequent releases of data can increase disclosure risks by serving as "auxiliary information" for an attacker in compromising the original release. Further, it is increasingly recognized that explicit interventions related to controls on downstream uses of data may be necessary to mitigate harm to data subjects, so consideration of privacy risks and controls on subsequent use is necessary.

---

320. See Craig Gentry, *Fully Homomorphic Encryption Using Ideal Lattices*, 41 ACM SYMPOSIUM ON THEORY OF COMPUTING 169 (2009).

321. See, e.g., O'HARA, *supra* note 265.

322. See Conley et al., *supra* note 19, at 843–44.

Transparency and accountability for misuse are essential to achieving an optimal balance of social benefit and individual privacy protection.<sup>323</sup> In the face of ubiquitous data collection practices, individuals find it difficult to effectively withhold consent because the playing field is uneven,<sup>324</sup> making accountability for misuse of increasing importance. In addition, data collectors and individual subjects of the data generally must be better informed of potential and actual uses of data. One tool for achieving this type of transparency for data subjects is a privacy dashboard that provides notice to individuals regarding which entities are accessing their data, how they are using the data, and any privacy risks they may be exposed to as a result of the use of their data.<sup>325</sup> Non-governmental organizations, privacy commissioners, and the public should be able to monitor government releases of data and speak out about privacy violations.<sup>326</sup> Accountability for misuse includes enabling individuals to find out how their data have been shared and used, civil and criminal penalties for privacy violations, and private rights of action for individuals harmed by an improper use of their data.

Organizations making data available online often provide terms of service or refer to ethical codes that describe guidelines and best practices for using confidential data about individuals. When sharing data in an individual transaction with a third party, data use agreements are a common approach to controlling use, sharing, and reuse. Laws or institutional policies may require data use agreements as a precondition for transferring certain types of sensitive information. Laws and policies sometimes specify the terms that must be included or the procedures that must be followed in drafting such an agreement,<sup>327</sup> or an institution may adopt a model contract that mirrors regulatory requirements and best practices within an industry. Data use agreements typically address limitations on use, sharing, and reuse of the data; obligations to secure the data; liability for harm arising from use or misuse of the data; and mechanisms for enforcing the terms of the agreement. In practice, it is

---

323. See Daniel J. Weitzner et al., *Information Accountability*, 51 COMM. ACM 82 (2008).

324. See Paul M. Schwartz & Daniel Solove, *Notice and Choice: Implications for Digital Marketing to Youth*, in SECOND NPLAN/BMSG MEETING ON DIGITAL MEDIA AND MARKETING TO CHILDREN (2009).

325. See, e.g., Molinari & Marsh, *supra* note 274, at 313–14; van der Sloot, *supra* note 268, at 149.

326. See, e.g., Keenan, *supra* note 279.

327. See, e.g., Health Insurance Portability and Accountability Act, 45 C.F.R. § 164.514(e) (2014) (providing the required terms to be included in data use agreements for sharing limited datasets).

often difficult to detect violations of a data use agreement and to enforce its terms; moreover, it is administratively costly to draft a data use agreement that is specific to the types of data and the actors involved in a given transaction,<sup>328</sup> though there have been recent proposals to automate the generation of custom data use agreements.<sup>329</sup>

Audit systems include both legal and technical mechanisms for detecting misuse of information and preventing individuals from violating a data use policy. A secure data enclave may be used to record every interaction with the data in an immutable audit log that can be reviewed later for improper uses of the data.<sup>330</sup> Such systems require users to register and provide contact information, and, in the event of discovery of disclosure risks in a given set of data, administrators can use audit logs to identify individuals who have previously accessed the data and request that they return or destroy the compromised information. Third party audits may be required to review data privacy and security procedures on an annual basis to ensure they are adequate, and such audits may also be required for contractors with access to the data.

The combination of lifecycle phase and means of control forms a grid (illustrated in Table 1) that can be used to identify feasible sets of controls based on policymakers' capabilities and scope of action. As noted below, some controls are applicable across multiple stages. Further, as described in Section III.D, one can select, from among these feasible controls, appropriate tools for minimizing the threats and vulnerabilities at each stage subject to the desired uses and expected benefits of the data.

---

328. See O'HARA, *supra* note 265, at 109.

329. Examples include a National Cancer Institute Center for Bioinformatics and Information Technology initiative to develop a tool for creating standardized electronic data use agreements, and research by members of the Privacy Tools for Sharing Research Data project at Harvard University and MIT exploring theoretical frameworks that could support development of an automated and modular data use agreement generator.

330. See Jonas & Harper, *supra* note 262.

Table 1: Example categorization of privacy controls and interventions.

	Procedural	Economic	Educational	Legal	Technical
Collection/ Acceptance	<ul style="list-style-type: none"> <li>• Collection Limitation Data</li> <li>• Data Minimization</li> <li>• Data Protection Officer</li> <li>• Institutional Review Boards</li> <li>• Notice and Consent Procedures</li> <li>• Purpose Specification</li> <li>• Privacy Impact Assessments</li> </ul>	<ul style="list-style-type: none"> <li>• Collection Fees</li> <li>• Markets for Personal Data</li> <li>• Property Rights Assignment</li> </ul>	<ul style="list-style-type: none"> <li>• Consent Education</li> <li>• Transparency</li> <li>• Notice</li> <li>• Nutrition Labels</li> <li>• Public Education</li> <li>• Privacy Icons</li> </ul>	<ul style="list-style-type: none"> <li>• Data Minimization</li> <li>• Notice and Consent</li> <li>• Purpose Specification</li> </ul>	<ul style="list-style-type: none"> <li>• Computable Policy</li> </ul>
Transformation	<ul style="list-style-type: none"> <li>• Process for Correction</li> </ul>		<ul style="list-style-type: none"> <li>• Metadata</li> <li>• Transparency</li> </ul>	<ul style="list-style-type: none"> <li>• Right To Correct or Amend</li> <li>• Safe Harbor De-Identification Standards</li> </ul>	<ul style="list-style-type: none"> <li>• Aggregate Statistics</li> <li>• Computable Policy</li> <li>• Contingency Tables</li> <li>• Data Visualizations</li> <li>• Differentially Private Data Summaries</li> <li>• Redaction</li> <li>• SDL Techniques</li> <li>• Synthetic Data</li> </ul>
Retention	<ul style="list-style-type: none"> <li>• Audits</li> <li>• Controlled Backups</li> <li>• Purpose Specification</li> <li>• Security Assessments</li> <li>• Tethering</li> </ul>		<ul style="list-style-type: none"> <li>• Data Asset Registers</li> <li>• Notice</li> <li>• Transparency</li> </ul>	<ul style="list-style-type: none"> <li>• Breach Reporting Requirements</li> <li>• Data Retention and Destruction Requirements</li> <li>• Integrity and Accuracy Requirements</li> </ul>	<ul style="list-style-type: none"> <li>• Computable Policy</li> <li>• Encryption</li> <li>• Key Management (and Secret Sharing)</li> <li>• Federated Databases</li> <li>• Personal Data Stores</li> </ul>
Access/ Release	<ul style="list-style-type: none"> <li>• Access Controls</li> <li>• Consent</li> <li>• Expert Panels</li> <li>• Individual Privacy Settings</li> <li>• Presumption of Openness vs. Privacy</li> <li>• Purpose Specification</li> <li>• Registration</li> <li>• Use Restrictions</li> <li>• Risk Assessments</li> </ul>	<ul style="list-style-type: none"> <li>• Access/Use Fees (for Data Controller or Subjects)</li> <li>• Property Rights Assignment</li> </ul>	<ul style="list-style-type: none"> <li>• Data Asset Registers</li> <li>• Notice</li> <li>• Transparency</li> </ul>	<ul style="list-style-type: none"> <li>• Integrity and Accuracy Requirements</li> <li>• Data Use Agreements (Contract with Data Recipient)</li> <li>• Terms of Service</li> </ul>	<ul style="list-style-type: none"> <li>• Authentication</li> <li>• Computable Policy</li> <li>• Differential Privacy</li> <li>• Encryption (Incl. Functional, Homomorphic)</li> <li>• Interactive Query Systems</li> <li>• Secure Multiparty Computation</li> </ul>
Post-Access (Audit, Review)	<ul style="list-style-type: none"> <li>• Audit Procedures</li> <li>• Ethical Codes</li> <li>• Tethering</li> </ul>	<ul style="list-style-type: none"> <li>• Fines</li> </ul>	<ul style="list-style-type: none"> <li>• Privacy Dashboard</li> <li>• Transparency</li> </ul>	<ul style="list-style-type: none"> <li>• Civil/ Criminal Penalties</li> <li>• Data Use Agreements</li> <li>• Terms of Service</li> <li>• Private Right of Action</li> </ul>	<ul style="list-style-type: none"> <li>• Computable Policy</li> <li>• Immutable Audit Logs</li> <li>• Personal Data Stores</li> </ul>

C. IDENTIFYING INFORMATION USES, THREATS, AND  
VULNERABILITIES

Assessing and treating privacy risk should address the range of threats to privacy; the vulnerabilities that exacerbate those threats; the likelihood of disclosure of information given those threats and vulnerabilities; and the extent, severity, and likelihood of harms arising from those disclosures.<sup>331</sup> This Section discusses examples of intended uses, threats, and vulnerabilities that should be considered in such an analysis.

1. *Information Uses and Expected Utility*

Selection of privacy controls should take into account the information uses and the utility of the data. Much of this analysis comes into play at the release stage, but use may occur at each stage of the lifecycle. Identifying the information uses involves a consideration of the uses intended by the legislators, regulators, and judges who established the relevant data collection, maintenance, and release policies; by the government agencies implementing the data programs; by the data subjects who provided their data to the government; by the data users who seek to access and analyze the data; and by the general public, or its expectations regarding how data about citizens are collected, retained, used, and released by the government. In addition, this analysis takes into account the stakeholders to whom benefits of the data program accrue, and the assumptions under which the benefits are expected to be realized.

Evaluating the utility of the data involves a comparison of the types of uses or analytic purposes intended by each of the stakeholder groups, and how the privacy controls at each stage enable or restrict such uses. The choice of a data release mechanism can enable or preclude different types of data uses. The organizations releasing data, and analysts who seek to use the data, may have certain uses in mind, such as requirements for conducting individual-level vs. population-level analyses, linking the released information with other data sources, or analyzing static sets or streaming, real-time data. A data release decision affects the output of the data, such as whether the data are made available as raw individual-level data, as a summary table, as model parameters, or as a static or dynamic visualization, among other alternatives. Similarly, the type of analysis desired by the analyst can vary between contingency tables, summary statistics, regression models, data mining, and other analysis types. For instance, a release of data by the U.S. Bureau of Labor Statistics contains

---

331. See Vadhan et al., *supra* note 24.

only aggregate-level data to enable statistical analyses at the population level, but not learning about individual respondents in the data.<sup>332</sup> In contrast, a data release under a state public records law, such as a response to a request for a list of handgun permit holders<sup>333</sup> or political donors,<sup>334</sup> will sometimes disclose information at the level of an individual. In the latter case, the release of such data at the individual-level may be appropriate if it is deemed to be vital to serving a public interest, such as enabling journalists and researchers to study the impact of handgun permitting on gun violence or to investigate the funding sources for a political campaign, respectively.

Consider, for example, the recent disclosure of automated license plate reader data by the City of Minneapolis.<sup>335</sup> These records were originally collected by local law enforcement officials for internal use in law enforcement investigations. The state legislature had recently passed an open data statute mandating the disclosure, in response to a request from the public, of all government data not specifically barred from release by a federal or state law or by a temporary classification of the data as nonpublic data.<sup>336</sup> As required by law, the city's police department released at least 2.1 million license plate reader records on file including the date, time, and location of automobiles throughout the city.<sup>337</sup> These data were used by commercial entities, such as vehicle repossession businesses and data aggregation services, in ways that were not intended by the legislature or the police department and that were inconsistent with public expectations about the uses of personal data collected and held by the government. News stories about the scope of data released and how they were being used by third parties led to public outcry about potential privacy violations, and the license plate reader data were soon after reclassified by the city as nonpublic records.<sup>338</sup> The intended law

---

332. See, e.g., Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) § 512(b)(1), 44 U.S.C. § 3501 note (2012).

333. See Fitz-Gibbon, *supra* note 78.

334. See *ProtectMarriage.com v. Bowen*, 752 F.3d 827, 835 (9th Cir. 2014).

335. See Eric Roper, *August 17, 2012: City Cameras Track Anyone, Even Minneapolis Mayor Rybak*, STAR TRIBUNE (Sept. 19, 2014), <http://www.startribune.com/aug-17-2012-city-cameras-track-anyone-even-minneapolis-mayor-rybak/166494646>.

336. See Minn. Stat. § 13.03 (2012).

337. See Cyrus Farivar, *Found: Secret Location of Minneapolis Police License Plate Readers*, ARSTECHNICA (Dec. 18, 2012), <http://arstechnica.com/tech-policy/2012/12/found-secret-location-of-minneapolis-police-license-plate-readers>.

338. See Minnesota Department of Administration, Information and Analysis Division, *Current Temporary Classifications*, <http://www.ipad.state.mn.us/docs/tccurrent.html> (last visited June 30, 2015).

enforcement use of the data and the public safety purpose of the data collection were not furthered by making the data available to the public. Instead, the stakeholders who benefited the most from the release were commercial entities who derived financial gain from use of the individual-level data that would not have been possible with aggregate data. Moreover, the transparency aims of the open data law could largely be served by the release of summary data, rather than individual-level data, about automated license plate reader programs. For these reasons, it is clear that the data release was not well-matched to the intent, expectations, and desired uses of the stakeholders involved.

## 2. *Privacy Threats*

A privacy analysis should explicitly consider the privacy threats, or the potential adverse circumstances or events that could harm a data subject as a result of the inclusion of that subject's data in a data collection, storage, management, or release. The concept of a privacy threat encompasses the capabilities and goals of adversaries and the sensitivity of the information, or its overall potential to cause individual, group, or social harm. Characterizing the types of threats to a data release and the types of harms that may result from the realization of those threats is a useful first step in estimating the extent of the potential adverse effects of a data release.

In some cases, characterizing the types of potential harms may put upper bounds on the overall expected harm associated with a release: for example, the only conceivable harm might be embarrassment. However, in other cases, evaluating the extent and severity of potential harm requires specifying an implicit or explicit threat model, a concept derived from the field of computer science.<sup>339</sup> Following such an approach, we aim to comment on additional desiderata for applying threat models within a lifecycle framework. Within information security it is a relatively standard practice to characterize the origin of threats using three broad categories: environmental, accidental, and deliberate acts.<sup>340</sup> Most discussions of information privacy issues related to data releases appear to be concerned entirely with deliberate privacy violations. However, when conducting a lifecycle analysis, one should also consider threats due to accident (e.g., software defects or mistaken release of data); such events are considered a significant risk in data management.<sup>341</sup> Privacy threats of environmental

---

339. For a general detailed and thoughtful discussion of threat models in the privacy context, see Wu, *supra* note 23, and Rubinstein & Hartzog, *supra* note 256.

340. See NIST, *CONTROLS* (draft), *supra* note 189.

341. See, e.g., Stephen Ohlemacher, *Census Bureau Admits Privacy Breach*, USA TODAY (Mar. 7, 2007), <http://usatoday30.usatoday.com/news/washington/2007-03-07>

origin (e.g., a system failure caused by equipment overheating) are conceivable, but unlikely.

When the origin of a threat is deliberate, a threat model can be thought of as an adversary model. Modeling adversaries typically includes specifying their objectives, the auxiliary knowledge they possess, and their resources or capabilities. Some broad examples of potential adversaries include nosy neighbors (or relatives), business competitors, data brokers, muckraking journalists, former spouses, potential employers or insurers, oppressive governments, and countless others. For example, a nosy neighbor might be characterized as having an objective to learn specific sensitive characteristics about a few particular subjects, detailed auxiliary information on these subjects, but few additional resources, whereas a data broker might be characterized as having a more general goal to link at least one person to a known record in the database, little general knowledge, but moderate resources.<sup>342</sup>

Some formal definitions of identifiability embed adversary models. For example, indistinguishability-based approaches such as  $k$ -anonymity imply that adversaries do not possess auxiliary knowledge of subject characteristics contained in the data, other than those characteristics labeled “quasi-identifiers,” whereas differential privacy assumes no limits on auxiliary adversary knowledge. However, neither  $k$ -anonymity nor differential privacy is designed to reduce harms from system vulnerabilities. One should keep the limitations of such implicit threat models in mind when performing a lifecycle analysis. In particular, since the information lifecycle generally involves retention of data, threat models that focus only on release are necessarily incomplete. For example, applying  $k$ -anonymity or other de-identification techniques to data before release may mitigate the threat of reidentification attacks against published data, but the technique is not designed to mitigate threats to privacy from observation of the data collection process, attacks against the servers that store the original data after it is collected, or post-publication releases of additional data that expand the auxiliary information available to an adversary.

One should also consider the sensitivity of the data, or the extent, type, and likelihood of harms that could result when a threat is realized. Generally, information should be treated as sensitive when that information, if linked to a person, even partially or probabilistically,

---

-1535966293\_x.htm (reporting that the Census Bureau inadvertently posted personal information publicly while testing new software).

342. See WILLENBORG & DE WAAL, *supra* note 136.

possibly in conjunction with other information, is likely to cause significant harm to an individual, group, or society. For instance, harms may occur directly as the result of a reaction of a data subject or third parties to the information, or indirectly as a result of inferences made from information. As an example of a potential harm that is indirect and inferential but nevertheless substantial, researchers have demonstrated that Facebook “likes” can be used to “automatically and accurately predict a range of highly sensitive personal attributes including sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender.”<sup>343</sup> A released set of data may, therefore, be very sensitive and have the potential to cause serious harm, even if it does not contain pieces of information that have traditionally been considered sensitive.

There is a broad range of informational harms recognized by regulation and by researchers and practitioners in the behavioral, medical, and social science fields.<sup>344</sup> Potential informational harms are wide ranging, including loss of insurability, loss of employability, market discrimination, criminal liability, psychological harm, loss of reputation, emotional harm, and loss of dignity. Broader harms to groups and society include social harms to vulnerable groups such as stereotyping or price discrimination; market failures (e.g., by enabling manipulation, or eliminating uncertainties on which insurance markets are predicated); and broad social harms arising from surveillance such as the chilling of speech and action, potential for political discrimination, or blackmail and other abuses.<sup>345</sup> In evaluating the sensitivity of information, it is also important to take into account the expected magnitude of the harm if identification or learning were to occur, and the number of people that would be exposed to harm if a privacy threat is realized.

### 3. *Privacy Vulnerabilities*

Recall from Section III.A that the definition of privacy vulnerabilities is broader than the corresponding information security term. In particular, privacy vulnerabilities are defined as any characteristics of the data,

---

343. Michal Kosinski et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. NAT'L ACAD. SCI. 5802 (2013).

344. See, e.g., ELIZABETH A. BANKERT & ROBERT J. ANDUR, INSTITUTIONAL REVIEW BOARD: MANAGEMENT AND FUNCTION (2006); RAYMOND M. LEE, DOING RESEARCH ON SENSITIVE TOPICS (1993).

345. See Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934 (2013); Daniel J. Solove, *A Brief History of Information Privacy Law*, in PROSKAUER ON PRIVACY (Christopher Wolf ed., 2006).

systems, and related context that increase the likelihood that privacy threats will be realized. Privacy vulnerabilities may arise from the characteristics of the data being collected, managed, or released; of the logical or physical systems used to manage that data; or of the broader context of release.

More specifically, vulnerabilities are associated with the scope of information collected, maintained, used, and disseminated by the organization. Some data programs involve the collection of a small number of data points about the characteristics of citizens in relation to a narrow topic or a single event. In other cases, governments have the potential to collect extensive (sometimes exhaustive), fine grained, continuous, and identifiable records of a person's location, movement history, associations, and interactions with others, behavior, speed, communications, physical and medical conditions, and commercial transactions, among many other categories of information. The choice of appropriate data sharing mechanism and privacy interventions will therefore differ for a police department periodically releasing crime statistics aggregated to the neighborhood level, and for an open data portal managing thousands of datasets containing a wide variety of geolocation, demographic, and survey response data.

Privacy vulnerabilities also arise from characteristics of the data being collected, managed, and released that make it easier to learn about the characteristics of individual data subjects. This set of characteristics can be thought of informally as the "identifiability" of the data. For example, there are risks that sensitive information about an individual will be disclosed through identity disclosure, meaning the risk of assigning a named individual to a sensitive record in a released set of data, as well as risks of attribute disclosure, meaning the risk of assigning a sensitive characteristic to an individual or group of individuals with or without associating this characteristic with a named individual. Attribute disclosure may occur, for instance, if an individual is known to be a member of a subsample in the data, and all members of that subsample share the same characteristic.

A traditional and commonly adopted approach to assessing disclosure risks begins by determining whether the data contain direct identifiers or quasi-identifiers, the latter of which are defined as personally identifiable, and externally readily observable, characteristics of individuals.<sup>346</sup> In the late 1990s, Latanya Sweeney identified the record of Massachusetts

---

346. See Alan F. Karr & Jerome P. Reiter, *supra* note 282.

Governor William Weld in an anonymized medical claims dataset by comparing sex, zip code, and date of birth with publicly available voter registration records.<sup>347</sup> These three seemingly innocuous pieces of information uniquely identify well over 50% of the U.S. population.<sup>348</sup> To mitigate the risk of identity disclosure, organizations typically make efforts to de-identify data by redacting direct identifiers, such as names, dates of birth, street addresses, telephone numbers, and Social Security numbers, and quasi-identifiers, such as sex, race, ethnicity, and other demographic information, before release. This is an approach that has historically been endorsed by laws and regulations in certain sectors. Health records redacted according to the HIPAA Privacy Rule safe harbor standard and education records redacted according to the FERPA de-identification standard can be shared without restriction because they are deemed not to contain identifiable information about individuals and therefore their release is considered minimally harmful.<sup>349</sup>

It is now well-understood, however, that stripping direct and indirect identifiers provides very weak privacy protections, as it is often quite easy to reidentify individuals in data that have been treated in this way.<sup>350</sup> It has been shown more generally that it takes very little information to uniquely identify an individual.<sup>351</sup> Even in the absence of direct identifiers and quasi-identifiers, disclosure risks can remain through indirect linkages to auxiliary information, or through statistical reidentification, through learning about individuals without identifying them (e.g., “attribute disclosure”), or through learning about characteristics of specific groups.<sup>352</sup> For instance, researchers demonstrated that individuals could be uniquely identified in a dataset containing “anonymized” film ratings by Netflix users, potentially allowing an individual’s religious, political, and sexual preferences to be inferred.<sup>353</sup> There have been numerous other examples

---

347. See Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J.L. MED. & ETHICS 98; Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* (Carnegie Mellon University, Data Privacy Working Paper 3, 2000), <http://dataprivacylab.org/projects/identifiability/paper1.pdf>.

348. See sources cited *supra* note 347.

349. See, e.g., 45 C.F.R. § 164.514(b) (2014); 34 C.F.R. § 99.31(b) (2014).

350. See, e.g., Ohm, *supra* note 22.

351. See de Montjoye et al., *Unique in the Shopping Mall*, *supra* note 225; Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 NATURE SCI. REP. 1376 (2013).

352. See sources at *supra* note 351.

353. See Narayanan & Shmatikov, *supra* note 224.

where this phenomenon has been exploited for reidentification,<sup>354</sup> and disclosure risks continue to grow as information about individuals is increasingly made available through publicly accessible government and commercial databases.<sup>355</sup>

More generally, the computational and statistical literature on privacy defines disclosure in a variety of ways. Work on statistical disclosure limitation initially defined disclosure, or risk of reidentification, operationally in terms of record linkage.<sup>356</sup> A record linkage occurs when a real person is matched with certainty to a specific record in the database. The use of record linkage as an operational definition for identifiability began to be generalized to concepts based on indistinguishability, following Latanya Sweeney's formalization of the concept of  $k$ -anonymity.<sup>357</sup> Indistinguishability can be thought of as hiding in the crowd, as each record in the database must be identical to some number of others on specified quasi-identifying fields. Most recently, disclosure has been defined in terms of learning. Formal privacy concepts such as differential privacy aim to place bounds on what one can learn from a particular release about any individual, as a result of her inclusion in the data from which the release was derived. We adopt this more modern definition.

To mitigate these types of attribute disclosure risks, some organizations go beyond redaction and also apply statistical disclosure limitation techniques to aggregate and perturb data before release.<sup>358</sup> However, aggregate data are also associated with disclosure risks. Providing query access to only aggregate statistics, for example, may reduce the risk of direct reidentification, but even such systems, if not carefully designed, can leak substantial amounts of personal information. For example, the Israel Central Bureau of Statistics provided a public web-based mechanism for people to make aggregate statistical queries of data from an anonymized survey, but researchers extracted the records of more

---

354. See, e.g., Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), <http://www.nytimes.com/2006/08/09/technology/09aol.html>.

355. See U.S. GOVERNMENT ACCOUNTABILITY OFFICE, INFORMATION RESELLERS: CONSUMER PRIVACY FRAMEWORK NEEDS TO REFLECT CHANGES IN TECHNOLOGY AND THE MARKETPLACE (2013), <http://www.gao.gov/assets/660/658151.pdf>.

356. See, e.g., Josep Domingo-Ferrer & Vicenç Torra, *Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage*, 13 STAT. & COMPUTING 343 (2003).

357. See Sweeney, *supra* note 224.

358. See Karr & Reiter, *supra* note 282.

than one thousand individuals by querying the system and, furthermore, demonstrated that it was possible to link the records to identifiable people.<sup>359</sup> It has also been demonstrated that a large number of aggregate genomic statistics could be used to determine, with high statistical confidence, whether an individual was part of the population studied, and this led the National Institutes of Health to eliminate public access to such statistics.<sup>360</sup> Researchers have even discovered attribute disclosure risks in recommendation systems such as Amazon's system for providing product suggestions based on aggregated consumer behavior.<sup>361</sup>

Addressing disclosure risks is a particularly challenging problem for high-dimensional datasets (i.e., datasets containing many attributes per individual), due to the quantity and richness of the data they contain.<sup>362</sup> For example, no method currently exists that allows detailed location data to be anonymized and then safely published. Rather, it has been demonstrated that individual mobility traces in a large-scale dataset of 1.5 million people are highly identifiable, with just four spatio-temporal points being sufficient to uniquely identify 95% of data subjects, and that coarsening such data provides very minimal privacy protection.<sup>363</sup> In another demonstration based on the credit card purchase histories for 1.1 million people, information about just four transactions was shown to be uniquely identifying for 90% of individuals.<sup>364</sup>

#### D. DESIGNING DATA RELEASES BY ALIGNING USE, THREATS, AND VULNERABILITIES WITH CONTROLS

In this Part, we have described the elements of a framework for assessing and managing privacy in data releases. The objective of the framework is to support the design of a program for data collection, management, and release that enables desired uses of the data and optimizes privacy and utility through selection of controls that are appropriate given the uses, threats, and vulnerabilities. In other words, a framework should map uses, threats, and vulnerabilities to privacy

---

359. See Ziv, *supra* note 20.

360. See Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLOS GENETICS 8 (2008); Jason Felch, *DNA Databases Blocked from the Public*, L.A. TIMES (Aug. 29, 2008), <http://articles.latimes.com/2008/aug/29/local/me-dna29>.

361. See Joseph A. Calandrino et al., "You Might Also Like." *Privacy Risks of Collaborative Filtering*, in PROCEEDINGS OF THE IEEE SYMPOSIUM ON SECURITY AND PRIVACY 231 (2011).

362. See NAT'L RESEARCH COUNCIL, PUTTING PEOPLE ON THE MAP, *supra* note 17.

363. See de Montjoye et al., *Unique in the Crowd*, *supra* note 351.

364. See de Montjoye et al., *Unique in the Shopping Mall*, *supra* note 225.

controls. In this Section we sketch this mapping in a broad outline. No single approach from privacy science, information science, computer science, or public policy is complete enough for a mapping to be fully prescriptive. Thus, this Section is intended to describe a systematic method for analyzing data release cases, not to determine specific outcomes.

A systematic approach to analysis, analogous to that used in information security, but adapted to the privacy arena, comprises specifying desired data uses and expected benefits; examining each stage of the cradle-to-grave data lifecycle to identify threats and vulnerabilities to privacy; and then selecting controls for each lifecycle stage that are consistent with the uses, threats, and vulnerabilities at that stage. We propose that a systematic analysis of privacy for data release include the elements that follow below.

We expect that, in the future, as emerging new privacy technologies become standardized and mature, and as the new privacy risks from big data became better understood, it will become possible to select controls for many common cases through a step-by-step engineering process. However, the state of the art does not yet support such a mechanical process for selecting interventions. Our aim is instead to provide a systematic and useful decomposition of the factors relevant to releasing data, in order to identify feasible interventions, manage privacy risks, and document decisions and the rationales supporting them. Because the selection of appropriate interventions depends on a specific evaluation of risks and benefits and there is not yet a standard mechanical process for this, we strongly recommend that government actors be transparent in documenting their analysis of each lifecycle stage and of the interventions selected.

### *1. Specifying Desired Data Uses and Expected Benefits*

It is a general truism that one should have some idea of the expected benefits of a policy before adopting it, and that the expected benefits should outweigh the costs and risks of the policy. In addition, as it has become clear that any government release of data about individuals creates some non-zero privacy risk, it is important to specify and articulate the expected benefits and the types of uses from which these expected benefits flow. Even where the expected benefits of a government data release are great (and we believe this is quite often the case), policymakers have an ethical responsibility to reduce risks to data subjects where possible. Government actors should thus select privacy controls that produce the smallest risks to data subjects possible while still realizing the expected benefits from the release.

Although the state of the art is not sufficiently mature to support precise recommendations or controls based on the analytic uses required, it is nevertheless useful to consider the analytic characteristics of the intended uses of the data, including the desired form of the analytic output; the goal of the analysis; the utility, loss, or quality measure; and the analysis methodology.<sup>365</sup> In addition, the compatibility of controls should be considered in light of the proposed analytic uses. For example, data minimization applied at the collection stage reduces the privacy risks to data subjects from both retention and release, but it can prevent many downstream uses that might be desirable. Functional encryption applied post-collection protects against threats during retention and allows for pre-specified families of queries to be performed over the data without revealing other information, although any uses that depend on richer queries than those for which the system was originally designed may be prevented. Providing differentially private analyses at the release stage can allow for statistical analysis of population-level properties, but cannot support analyses that target individuals or small subsets of the population. Applying redaction at the release stage and releasing an entire  $k$ -anonymized database for public use permits a wide variety of analytic models and derivative works to be produced, but the redaction necessary for privacy protection both reduces the utility in the data and potentially biases inferences based upon the redacted data.

## 2. *Selecting Controls*

As discussed in Section III.C, there is a range of threats to privacy and sources of vulnerability that make the threats more likely to manifest in a given set of data. The threats and vulnerabilities associated with a specific data release case vary according to the characteristics of the data, information systems, and actors involved, among other contextual factors. Such characteristics may exacerbate vulnerabilities, limit the types of privacy controls that can be feasibly applied, or reduce the effectiveness of such controls. As we have suggested in an earlier work,<sup>366</sup> the following data characteristics are particularly relevant to an analysis of vulnerabilities in a data release and the selection of appropriate controls:

---

365. See Alexandra Wood et al., *Integrating Approaches to Privacy Across the Research Lifecycle: Long-term Longitudinal Studies* (Berkman Ctr. for Internet & Soc'y, Research Pub. No. 2014-12, 2014), <http://ssrn.com/abstract=2469848>.

366. See *id.*

- Logical structure (e.g., single relation, multiple relational, network or graph, semi-structured, geospatial, and aggregate table);
- Source population and unit of observation or measurement;
- Attribute measurement type (e.g., continuous or discrete; ratio, interval, ordinal, or nominal scale; and associated schema or ontology);
- Performance characteristics (e.g., dimensionality or number of measures, number of observations or volume, sparseness, heterogeneity or variety, and frequency of updates or velocity); and
- Quality characteristics (e.g., measurement error, metadata, completeness, and total error).

For example, the characteristics associated with different forms of big data can have a variety of surprising privacy implications. Individual records in high-dimensional datasets (i.e., datasets containing many attributes per individual) are often unique, and thus it would be difficult to apply controls based on record linkage, such as  $k$ -anonymity.<sup>367</sup> Rich, messy data, such as information from social networks, can contain unanticipated information in the structure of the data that creates vulnerabilities, as the identifiability of the data will likely remain high after standard redaction controls have been applied.<sup>368</sup>

More generally, the degree of harm to be prevented should determine the resources that policymakers devote to privacy controls and interventions, and the extent to which barriers to use and reductions in data utility are justified. In turn, the expected harm from an uncontrolled release is a function of the threats and vulnerabilities from all stages of the information lifecycle. In many cases the primary threats and vulnerabilities arise from reidentification or learning vulnerabilities being realized after the data have been released, and the degree of harm can be roughly estimated by the category in which that harm falls. Once determined based on the threats and vulnerabilities of a release, the level of expected

---

367. See, e.g., Narayanan & Shmatikov, *supra* note 224.

368. See, e.g., Lars Backstrom et al., *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, in PROCEEDINGS OF THE 16TH INT'L CONFERENCE ON WORLD WIDE WEB 181 (2007); Jasmine Novak et al., *Anti-Aliasing on the Web*, in PROCEEDINGS OF THE 13TH INT'L CONFERENCE ON WORLD WIDE WEB 30 (2004).

harm from an uncontrolled release can help guide the selection of an appropriate set of privacy controls.

Figure 2 below illustrates a partial conceptualization of the relationship between the threats and vulnerabilities for a set of data and the suitability of selected procedural and legal controls implemented at the collection and release stages. For purposes of illustration, this diagram focuses on a small subset of interventions from the more comprehensive set of procedural, economic, educational, legal, and technical controls cataloged in Section III.B. In practice, the design of a data release mechanism should draw from the wide range of available interventions and incorporate controls at each stage of the lifecycle, including the post-access stage.

In this diagram, the  $x$ -axis provides a scale for the level of expected harm from uncontrolled use of the data, meaning the maximum harm the release could cause to some individual in the data based solely on the sensitivity of the information (i.e., the use of a privacy control is not a factor in the calculation of the level of expected harm). This scale ranges from low to high levels of expected harm, with harm defined to capture the magnitude and duration of the impact a misuse of the data would have on an affected individual's life. To illustrate how such a scale could be used, we have placed a number of examples as reference points along this axis. At one end of the axis, there are the most negligible harms, or those that are not expected to have an effect on an individual's daily life. At the other end, there are life-threatening harms, such as harms that may occur if data about domestic violence victims or individuals engaged in gang-related activity are leaked. Between these two endpoints fall examples of minor and temporary harms, significant and lasting harms, and life-altering harms that fall short of being life-threatening.

The  $y$ -axis provides a scale for the post-transformation identifiability or learning potential from a data release. In contrast to the level of expected harm, the assessment of information identifiability or learning potential may be affected by the application of a privacy control. A number of examples are provided along this scale for illustration purposes. At one end are datasets containing direct or indirect identifiers such as names, addresses, and dates of birth. At the other end are data released using expertly applied rigorous statistical and non-statistical disclosure limitation techniques, particularly those supported by a formal mathematical proof such as differential privacy or secure multiparty computation. In between, there are examples such as datasets from which direct and indirect identifiers have been removed and data transformed using heuristic statistical or non-statistical disclosure limitation

techniques, or those based on experience and intuition such as traditional aggregation techniques.

The level of expected harm from uncontrolled use and the post-transformation identifiability of the data, taken together, point to privacy controls that are appropriate in a given case, as shown by the shaded regions in the diagram. Regions divided by a diagonal line correspond to categories of information for which a government agency could reach different conclusions based on the intended uses of the data and privacy standards that vary based the applicability of a regulation, contract, institutional policy, or best practice.

The white region of the diagram represents categories of data that one might reasonably decide to release without the use of additional privacy controls such as terms of service restricting data uses. For example, the lower left corner of the diagram corresponds to information associated with negligible harm from uncontrolled use and information to which rigorous disclosure limitation techniques providing a formal privacy guarantee have been applied. This is a category of information one might reasonably decide to release without the use of additional privacy controls, unless such controls are required by regulation, contract, or policy (in which case, the policy controls required by such policy should be applied). For example, in many cases it would likely be considered reasonable to release certain differentially private statistics on basic demographics of a population, such as age distribution, without requiring additional restrictions on use or redisclosure.

Regions in light gray refer to data releases associated with a low level of expected harm from uncontrolled use, or data that have been transformed to reduce the identifiability of the data. For most data in this category, notice to and consent from the data subjects, in combination with clickthrough terms of service prohibiting misuses of the data, would be considered a reasonable practice for release. An example of data in this category is national and state test scores that were released as custom aggregate statistics by the Department of Education's National Center for Education Statistics under terms of service prohibiting reidentification and linking of the data, among other restrictions.<sup>369</sup> For some data within this category, such as data collected from human subjects for research conducted with federal funding, approval from an institutional review

---

369. See *NAEP Data Explorer*, NAT'L CTR. FOR EDUC. STATISTICS, <http://nces.ed.gov/nationsreportcard/naepdata> (last visited Aug. 10, 2015).

board and a data use agreement may be required, as reflected by space shared with the medium gray region.<sup>370</sup>

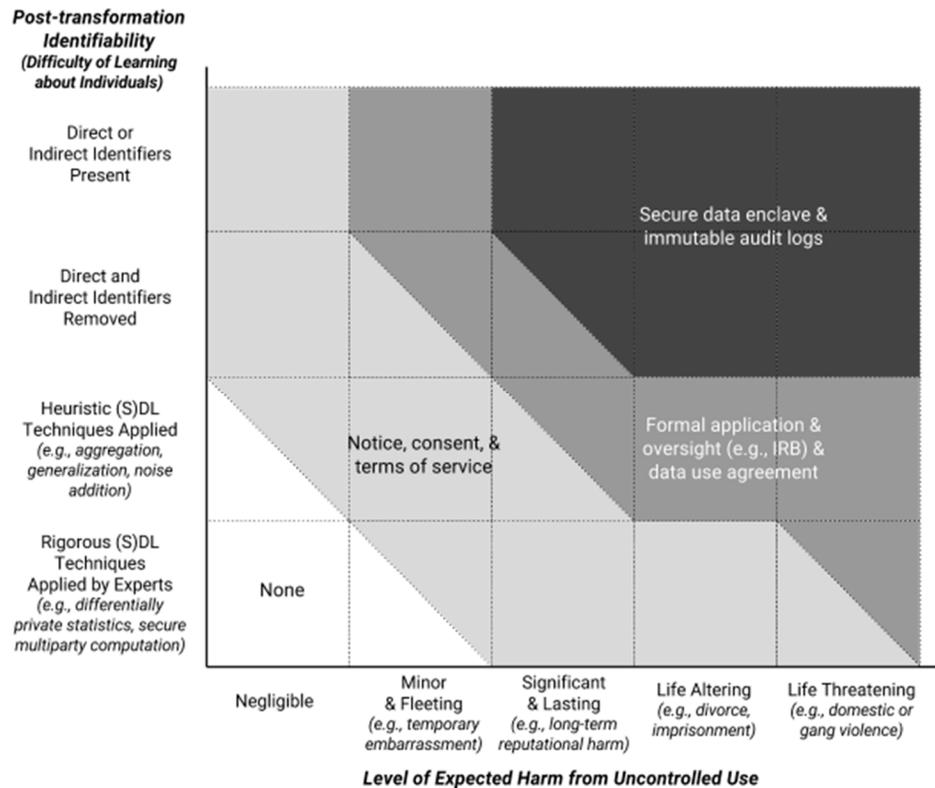


Figure 2: Conceptual diagram of the relationship between post-transformation identifiability, level of expected harm, and suitability of selected privacy controls for a data release.

The medium gray region corresponds to higher levels of expected harm or increased identifiability of the data. This category of data is released only upon application, review, and oversight from a data protection officer or institutional review board (IRB) under a data use agreement limiting data use and setting forth penalties for misuse. Examples of data within this category include certain medical or educational records protected by HIPAA and FERPA, respectively, which can be shared only in limited circumstances with screened individuals under the terms of a data use agreement or with IRB approval.<sup>371</sup>

370. See the Federal Policy for Human Subjects Research, 45 C.F.R. pt. 46 (2014).

371. See FERPA, 34 C.F.R. §§ 99.31(a)(6)(iii)(C), .35(a)(3) (2014); HIPAA Privacy Rule, 45 C.F.R. §§ 164.512(i)(1)(i), .514(2) (2014).

For highly identifiable and harmful data, represented by the darkest gray region, access may be reasonably permitted only through a secure data enclave with immutable audit logs and enforcement mechanisms. Examples of information in this category include responses to sensitive survey questions, such as those related to abortion, illegal conduct, sexual behavior, stigmatizing medical conditions, and mental health,<sup>372</sup> maintained by statistical agencies in identifiable form and therefore protected by CIPSEA.<sup>373</sup>

In some cases, the practices recommended by this diagram deviate from existing regulatory standards. Consider, for example, the region corresponding to data associated with significant and lasting harms but from which direct and indirect identifiers have been removed. An example of information from this category is medical records which would otherwise be protected by the HIPAA Privacy Rule but which have been transformed by redaction of certain direct and indirect identifiers according to the law's safe harbor de-identification standard or its limited dataset standard for de-identification. If redacted according to the safe harbor standard, the data can be released without any restriction, and, if redacted according to the limited dataset standard, the data can be shared under the terms of a data use agreement. Also note that the privacy science literature has called into question the effectiveness of simple redaction of direct and indirect identifiers for privacy protection.<sup>374</sup> In light of evolving best practices, an agency may decide not to adopt the safe harbor de-identification standard but to require privacy controls, such as application and oversight procedures, that are more restrictive than the law explicitly requires. Such an approach deviates from common practice, but it could be considered a best practice for an agency seeking to provide strong privacy protections in light of current understanding of disclosure risks.

The diagram illustrates that, for any data collected about individuals, there should at minimum be some terms of service restricting their use, unless the data are deemed negligibly harmful and they have been transformed to reduce disclosure risks. Figure 2 also illustrates how, for a given set of data, access may be made available to different categories of users through different modes of release, an approach referred to as a tiered access model. The diagram shows the relationship between

---

372. See U.S. Census Bureau Data Stewardship Executive Policy Committee, Policy DS-16, Policy on Respondent Identification and Sensitive Topics in Dependent Interviewing (Dec. 9, 2014), [http://www2.census.gov/foia/ds\\_policies/ds016.pdf](http://www2.census.gov/foia/ds_policies/ds016.pdf).

373. For a discussion of the use of formal screening processes and secure data enclaves for accessing statistical information, see *supra* Part II.A.3.

374. See discussion *supra* Section III.C.3.

transformation and release controls, and indicates how controls can be selected at each access tier. For example, an agency could provide public access to some data without restriction after robust disclosure limitation techniques have transformed the data into, for example, differentially private statistics. Data users who intend to perform analyses that require the full dataset, including direct and indirect identifiers, could be instructed to submit an application to an institutional review board or other oversight body, and their use of the data would be restricted by the terms of a data use agreement. In this way, data release mechanisms can be tailored to the threats and vulnerabilities associated with a given set of data, as well as the uses desired by different users.

Note that, although the data transformation and release stages typically attract the most attention, threats and vulnerabilities arising from other lifecycle stages should not be ignored. For example, privacy risks may be present at the collection stage if the data collection process could be observed by an adversary; data retained in long-term storage are vulnerable to unintended breaches; and, increasingly in a big data world, external, independent publication of auxiliary information may create new or unanticipated privacy risks long into the post-access stage.

#### **IV. APPLYING THE FRAMEWORK TO REAL-WORLD EXAMPLES OF GOVERNMENT DATA RELEASES**

To demonstrate how this analytical framework can inform the selection of privacy controls that align with the uses, threats, and vulnerabilities that are specific to a data release, this Part applies the framework to two real-world examples of open government data releases. The first examines a proposed rule from the Occupational Safety and Health Administration to make workplace illness and injury records publicly available in a searchable online database. The second analyzes the open data portals for Boston and Seattle and the policies that guide them. These real-world data release cases are used to illustrate, at a relatively fine level of detail, the types of uses, threats, vulnerabilities, and controls that should be considered by government agencies when collecting, retaining, transforming, and releasing data about individuals. In addition, this discussion describes how privacy controls and interventions can be matched to the uses, threats, and vulnerabilities associated with these data release cases. The data releases reviewed in this Part describe specific examples of data handling practices that are widespread, and gaps or misalignments identified below are representative of many of the types of issues that arise in government data releases rather than issues specific only to the cases discussed.

### A. PUBLIC RELEASE OF WORKPLACE INJURY RECORDS

The Occupational Safety and Health Administration (OSHA), a federal agency overseeing workplace health and safety conditions, requires companies in designated industries to create and maintain records on illnesses, injuries, and deaths that occur at their work sites. In an ongoing rulemaking initiated in 2013, OSHA has proposed expanding its collection of the illness and injury records maintained by these establishments and publishing the data via a searchable web interface.<sup>375</sup> To better understand the impact of OSHA's proposal for expanding the collection and release of data about workplace illness and injuries, this analysis examines the uses, threats, vulnerabilities, and controls at each stage of the lifecycle of the proposed program.

#### 1. *Collection and Acceptance Stage*

When collecting information about humans and human behaviors, a government agency should specify the intended uses of the data and the expected benefits of the program. The rationale underlying OSHA's proposed rule is that regular collection of workplace injury and illness data in electronic form will help OSHA compare illness and injury rates between establishments and thereby detect poor health and safety conditions. The proposed rule seeks to expand the collection of these data so that OSHA can release the data publicly, enabling employers and employees, members of the press, and researchers to examine the data and exert pressure on companies with poor health and safety records. Routinizing the collection and dissemination of the data is expected to bring economic gains, since it is well established that, when the cost of monitoring incidents is low, more regular monitoring and gradual sanctions increase social welfare benefits.<sup>376</sup> Furthermore, the costs of occupational injury in the United States are at minimum tens of billions of dollars annually.<sup>377</sup> Most of these costs are not borne by the firms in which injuries occur, or by insurers, but are instead imposed on the individual and on the societal safety net.<sup>378</sup> For these reasons, reductions in injury brought about through better detection and changes in individual and firm behavior have the potential to yield substantial benefits to individuals and

---

375. Proposed Rule, Improve Tracking of Workplace Injuries and Illnesses, 78 Fed. Reg. 67254 (Nov. 8, 2013).

376. See A. Mitchell Polinsky & Steven Shavell, *The Economic Theory of Public Enforcement of Law*, 38 J. ECON. LITERATURE 45 (2000).

377. See J. Paul Leigh, *Economic Burden of Occupational Injury and Illness in the United States*, 89 MILBANK Q. 728, 728–29 (2011).

378. See *id.*

to the economy. OSHA's choice of privacy controls should be tailored to these intended uses and enable comparisons of health and safety records at the establishment level if the expected benefits of the data collection program outweigh the attendant privacy risks.

When assessing the privacy risks associated with a data collection program, an agency should identify the privacy threats and vulnerabilities. The proposed rule would greatly expand the scope of information collected by OSHA. Currently, OSHA collects summary information such as the total number of illnesses and injuries at workplaces on an annual basis, and uses these data to calculate establishment-specific injury and illness rates.<sup>379</sup> The proposed rule would expand the scope of collection to include all incident-specific injury and illness records currently maintained by these companies.<sup>380</sup> Information such as names, addresses, and dates of birth would be removed before the records are reported to OSHA, but the records would include an employee's job title, the date of the injury or onset of illness, the location within the workplace where the injury occurred, a description of the injury or illness, a classification of the impact of the injury or illness, and the type of injury or illness.<sup>381</sup> Many examples from the reidentification literature illustrate how it is often possible to identify individuals in a database even after fields such as name, address, gender, and date of birth have been removed.<sup>382</sup> For example, some individual entries for a field, such as a job title held by only one person at a company or a description of an unusual injury, may be identifying on their own. In addition, although some of the information could be considered benign, there are situations in which details regarding an injury or illness may be sensitive. Recognizing the sensitivity of workplace injury and illness records, OSHA regulations currently provide additional protection for "privacy concern cases," which include a limited set of injuries or illnesses related to sexual assault, mental health, or infectious diseases.<sup>383</sup> However, there are additional types of injury or illness cases that may involve sensitive issues, such as drug and alcohol abuse, and the disclosure of this information could create substantial

---

379. *See* 78 Fed. Reg. at 67263.

380. *See id.*

381. *See id.* at 67259–60.

382. *See, e.g.,* Sweeney, *supra* note 20.

383. The privacy concern cases include "[a]n injury or illness to an intimate body part or the reproductive system; [a]n injury or illness resulting from a sexual assault; [m]ental illnesses; HIV infection, hepatitis, or tuberculosis; [n]eedlestick injuries and cuts from sharp objects that are contaminated with another person's blood or other potentially infectious material . . . ; and [o]ther illnesses, if the employee voluntarily requests that his or her name not be entered on the log." 29 C.F.R. § 1904.29(b)(7) (2014).

privacy risks and potential harms for the individuals involved. For these reasons, the information to be collected is likely sensitive and uniquely identifying for many of the individuals in the database, despite the privacy protections provided in the proposed rule.

The collection of individual incident records, and the uniqueness of such records, increase the risk that sensitive information about an individual will be disclosed in an intentional or unintentional breach as the records are collected. Moreover, it is not clear that the collection of detailed records about individual illness and injury incidents will substantially advance OSHA's aims to improve detection of inadequate health and safety practices compared to the collection of summary information about incidents at the establishment level. It is likely that establishments with poor health and safety records could be identified based on the total number of reported incidents over a period of time, and, for establishments with high numbers of incidents, OSHA could initiate an investigation to obtain additional details to determine whether an enforcement action should be brought against a specific establishment. In summary, the proposed rule calls for expanding the scope of potentially sensitive and identifiable information collected from an establishment, without a clear rationale for the intended uses and benefits of this additional information. These are indications that the agency should consider whether a privacy control at the point of collection, such as the implementation of a privacy risk assessment procedure, aggregation transformation, or data minimization principle, would be appropriate.

## *2. Retention Stage*

As noted, the proposed rule would greatly expand the scope of information reported to OSHA, and OSHA would retain this information within its databases. In addition to summary level information about the total number of illness and injury incidents at a given establishment, OSHA would retain detailed records related to each incident. This expansion of the scope of data retained by OSHA necessarily adds to the threats and vulnerabilities associated with the data. OSHA's retention of individual-level information from a vast number of establishments in a central repository increases the likelihood that the data would be the target of a hacker or that a large quantity of data would otherwise be disclosed in a data breach. In addition to considering whether the agency should adopt a principle of data minimization, OSHA should implement strong information security controls such as encryption, authentication, and audits of security practices, to protect the information

as it is held in storage. Although OSHA is subject to FISMA,<sup>384</sup> the proposed rule does not specify the FISMA risk level that would be assigned to the data or which information security controls would be implemented for the new categories of data to be collected and stored under this policy.

### 3. *Post-Retention Transformation*

The proposed rule would require the public release of all workplace illness and injury records collected by OSHA, and would not require OSHA to transform the data in any way prior to release. For instance, it would not require a pre-release review of the data for sensitive information or require any further redaction, aggregation, or recoding of values before the data are shared with the public. OSHA would not have to look far to find examples of review mechanisms, however, because OSHA regulations require employers to review and remove “personally identifying information” before sharing workplace injury and illness records with non-governmental or contracted third parties.<sup>385</sup> Outside of the limited set of privacy concern cases, which seem to be underinclusive of all privacy-sensitive incidents, employers are not directed by the regulations to systematically review and redact personally identifying information from incident descriptions, or to prevent private information from being easily inferred by such redactions. OSHA may not even be aware of the extent to which identifying information might be present in descriptive fields, given that it does not routinely access or collect the injury and illness reporting forms outside of the limited number of investigations and inspections the agency conducts each year.

### 4. *Release and Access Stage*

To identify the privacy vulnerabilities at the release and access stage, an agency should consider the scope of information covered. OSHA proposes to publish all workplace illness and injury records that are not barred from release by FOIA, the Privacy Act, or OSHA regulations.<sup>386</sup> OSHA interprets these laws as prohibiting the release of information such as name, address, date of birth, and gender, but not an employee’s job title, the date and time of an illness or injury incident, and descriptions of an

---

384. Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–49 (2012).

385. Specifically, employers must “remove or hide the employees’ names and other personally identifying information” before disclosing information on Forms 300 and 301 to third parties. 29 C.F.R. § 1904.29(b)(10).

386. *See* 78 Fed. Reg. at 67263.

injury or illness and where and how it occurred.<sup>387</sup> OSHA would therefore make both establishment-level and incident-level workplace injury and illness data from these records available online via a searchable database and in downloadable raw data files.<sup>388</sup> The searchable database, as proposed, would display tables containing information about each workplace such as the name, address, industry, total illness and injury case rates, and total employee days away.<sup>389</sup> It would also provide details for individual illness and injury incidents that occurred at large establishments, as shown in the mockup of the web interface in Figure 3.<sup>390</sup> Notably, the incident-level records would include a free-form text field describing the employee's activities at the time of the injury, the circumstances that contributed to the injury, and the extent of injury.

FOIA and the Privacy Act—the legal standards for privacy protection that are cited in the rulemaking—provide little guidance for gauging privacy risks, and it is not clear that these laws are suitable benchmarks for determining the scope of workplace illness and injury information that is appropriate for public disclosure. For instance, the Privacy Act applies only to systems of records that enable information to be retrieved by an individual's name or identifying number,<sup>391</sup> but OSHA's database would maintain records according to the establishment name, rather than an individual's name. FOIA is problematic as a standard because it is designed as a discretionary request-response system in which requests are individually reviewed for privacy risks, and it is not well-suited for a system in which unstructured information in free-form text fields is categorically released to the public without prior review.<sup>392</sup>

In addition, the uniqueness of the individual records to be released makes it likely that a friend, family member, colleague, prospective employer or insurer, or marketer could potentially use personal knowledge of an incident or details from a news article to reidentify an individual in

387. *See id.* at 67259–60.

388. *See id.* at 67263. A final rule is anticipated to be published in March 2016.

389. *See* U.S. OCCUPATIONAL SAFETY & HEALTH ADMINISTRATION, IMPROVE TRACKING OF WORKPLACE INJURIES AND ILLNESSES RULEMAKING: MOCKUP OF PROPOSED WEB DISPLAY OF SUBMITTED INJURY/ILLNESS DATA (Apr. 22, 2013), <https://www.osha.gov/recordkeeping/LDCsys-rulemaking-Search.pdf>.

390. *See id.*

391. 5 U.S.C. § 552a(a)(5).

392. Legal scholars have also raised a number of concerns regarding the privacy protections for individuals in FOIA. *See, e.g.*, Bloom, *supra* note 19; Lisa Chinai, *Picture Imperfect: Mug Shot Disclosures and the Freedom of Information Act*, 9 SETON HALL CIR. REV. 135 (2012); Evan M. Stone, *The Invasion of Privacy Act: The Disclosure of My Information in Your Government File*, 19 WIDENER L. REV. 345 (2013).

the OSHA database and uncover sensitive details about the extent of an individual's injury or illness, and the circumstances leading up to it. In fact, OSHA regulations recognize that descriptions of injuries and illnesses may be identifying and encourage employers to exercise discretion in describing injuries or illnesses in a sensitive "privacy concern" case if they "have a reasonable basis to believe that information describing the privacy concern case may be personally identifiable even though the employee's name has been omitted."<sup>393</sup> Despite recognizing that privacy risks can persist even in redacted records, OSHA does not provide any mechanisms for addressing such risks for the majority of records it proposes to release. This approach provides weaker protection compared to standards from federal regulations, such as CIPSEA and the HIPAA Privacy Rule.<sup>394</sup>

---

393. 29 C.F.R. § 1904.29(b)(9) (2014).

394. Although the proposed data disclosures are likely not governed by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) or HIPAA, it is worth noting that these laws rely on definitions of personally identifying information that are significantly more expansive than the approach from the rulemaking. CIPSEA guidance states that "confidential information refers to any *identifiable* information, regardless of whether direct identifiers such as name and/or address have been removed from the individual records." OFFICE OF MGMT. & BUDGET, *supra* note 128 at 8. In addition, the HIPAA Privacy Rule states that individually identifiable health information is information that "relates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual" and that "identifies the individual or with respect to which there is a reasonable basis to believe the information can be used to identify the individual." 45 C.F.R. § 160.103 (2014). Either of these standards would prohibit the release of information such as a job title or an injury or illness description, to use just the examples above, that could reasonably be tied to an individual.



**UNITED STATES  
DEPARTMENT OF LABOR**

A to Z Index | En Español | Contact Us | FAQs | About OSHA

OSHA
OSHA QuickTakes Newsletter RSS Feeds Print This Page Text Size

Occupational Safety & Health Administration
We Can Help
What's New | Offices

Home Workers Regulations Enforcement Data & Statistics Training Publications Newsroom Small Business
OSHA

## 2008 Establishment Incident Report (OSHA 301) Printer Friendly

Establishment Data For: COMPANY NAME INC  
SIC: 1234 - Type of Business  
NAICS: 123456 - Type of Business

*OSHA's Form 301*  
**Injury and Illness Incident Report**

**Was employee treated in an emergency room?**  
 Yes  
 No

**Was employee hospitalized overnight as an in-patient?**  
 Yes  
 No

Case number from the Log  (Transfer the case number from the Log after you record the case.)

Date of injury or illness     
Month Day Year

Time employee began work   AM  PM

Time of event   AM  PM  Check if time cannot be determined

**What was the employee doing just before the incident occurred?** Describe the activity, as well as the tools, equipment, or material the employee was using. Be specific. *Examples:* "climbing a ladder while carrying roofing materials"; "spraying chlorine from hand sprayer"; "daily computer key-entry."

Lifting boxes on shelves while restocking products.

**What Happened?** Tell us how the injury occurred. *Examples:* "When ladder slipped on wet floor, worker fell 20 feet"; "Worker was sprayed with chlorine when gasket broke during replacement"; "Worker developed soreness in wrist over time."

Worker developed sharp pains in back while lifting a particularly heavy box.

**What was the injury or illness?** Tell us the part of the body that was affected and how it was affected; be more specific than "hurt," "pain," or "sore." *Examples:* "strained back"; "chemical burn, hand"; "carpal tunnel syndrome."

Worker strained his back and noted considerable pain and limitation of movement.

**What object or substance directly harmed the employee?** *Examples:* "concrete floor"; "chlorine"; "radial arm saw." *If this question does not apply to the incident, leave it blank.*

Lifting heavy boxes.

**If the employee died, when did death occur?** Date of death     
Month Day Year

Feedback | Disclaimer

U.S. Department of Labor | Frances Perkins Building, 200 Constitution Ave., NW, Washington, DC 20210  
 www.dol.gov | Telephone: 1-866-4-USA-DOOL | TTY: 1-877-889-5627 | Contact Us

Figure 3: OSHA's mockup of proposed web display of workplace injury and illness reports.<sup>395</sup>

395. *Id.*

At the same time, the proposed rule calls for information to be withheld that, by itself, is unlikely to pose a heightened disclosure risk. For example, fields indicating whether an injury resulted in an overnight hospital stay or emergency room visit must be removed.<sup>396</sup> These fields are arguably less likely to be identifying or sensitive than other fields that would be released such as detailed textual descriptions of the injury and illness. In addition, these fields would also provide information about the severity of the injury that would be useful for analysis. Thus, the redaction reduces the utility of the released data for scientific and policy analysis, and suggests that the standard for classifying fields as identifying or non-identifying is arbitrary.

#### 5. *Post-Access Stage*

The proposed rule does not provide any safeguards for protecting information after its release. It does not propose restrictions—technical, legal, or otherwise—on how these records, which may contain uniquely identifying information, may be used.<sup>397</sup> Transparency about releases of personal information, restrictions on disclosure, and accountability for misuse are all essential to achieving an optimal balance of social benefit and individual privacy protection.<sup>398</sup> More specifically, OSHA should consider implementing accountability mechanisms to enable individuals to see where data describing them has been distributed and used, set forth penalties for misuse, and provide individuals with a right of action to seek redress for harms caused by the release and misuse of their personal information.

#### 6. *Aligning Uses, Threats, and Vulnerabilities with Controls*

The rulemaking proposes to protect the privacy of individuals whose information would be released by requiring employers to withhold identifiers such as names, addresses, dates of birth, and gender, from the records transferred to OSHA. As mentioned above, the complexity, detail, richness, and emerging uses for data such as those to be released by OSHA create significant uncertainties about the ability of traditional de-identification methods, such as simple redaction, to protect confidential

---

396. See 78 Fed. Reg. at 67260 (prohibiting the release of fields 1 through 9 from a standard OSHA form, where fields 8 and 9 refer to whether the employee was treated in an emergency room or hospitalized overnight as an in-patient, respectively).

397. For example, safeguards could be imposed by a system that restricts access to the most sensitive data to only trusted users through technical means coupled with legal contracts specifying additional conditions on use (e.g., re-sharing of data, publishing identifying information, etc.).

398. See Weitzner et al., *supra* note 323.

information. Despite these uncertainties, the rulemaking does not propose additional privacy controls, such as requiring the release of only aggregate records, the generalization of free-form text fields as categorical values, or the use of other more advanced techniques to transform the data to provide privacy protection. OSHA also has not provided a rationale for requiring the collection and release of individual-level information. Moreover, the proposed rule appears to lack mechanisms that would provide accountability for harm arising from misuse of disclosed data. For these reasons, the privacy controls proposed by OSHA do not seem to align well with the intended uses of, or the privacy risks associated with, the data it plans to collect, retain, and release to the public.

OSHA should consider additional privacy controls that are calibrated to the specific uses, threats, and vulnerabilities associated with the data. Generally, one size does not fit all, and tiered modes of access, including public access to privacy-protected data and vetted access to the full data collected, should be provided. Making workplace injury and illness records available while also providing stronger privacy protections for employees can be informed by a careful consideration and balancing of the sensitivity, learning potential, intended uses, and expected benefits of the data. Publishing workplace injury and illness data using multiple levels of access, with embedded review and accountability mechanisms, could bring gains in both privacy and utility if properly implemented.

For data made available to the public without significant restriction, a good practice is to ensure that the data release process and method cause no individual to incur more than a minimal risk of harm from the use of her data, even when the released data are combined with other data that may be reasonably available. On this end of the privacy-utility spectrum, the unrestricted public release of data might be limited to aggregate information. Such a release could be similar in detail to the aggregate information currently provided by OSHA but include all of the firms that would be required to submit records under the proposed rule. Many members of the public would likely find that a series of contingency tables and visualizations could simplify their review and comparison of the workplace safety records of various employers. Within such aggregated releases, generalizing or coding open-ended fields such as injury and illness descriptions could additionally reduce the risk that sensitive details about an individual's injury or illness will be revealed. Further, it may be possible to release these kinds of aggregate statistics with both formal guarantees of privacy and accuracy using existing differentially private

methods.<sup>399</sup> Since large companies will likely have large numbers of incidents, adding noise to the statistics would likely not reduce their accuracy by very much.

To enable interactive analysis of the data, an intermediate level of access could be set up through a privacy-aware model server. This server would ensure that the results provided by the analysis leak minimal private information. It could also be used to permit audits of access and to impose some click-through data use agreements providing individuals with additional legal protections from misuse.

At the same time, for a user to gain the full utility of the data, she must have rich access to information that is minimally redacted and at the finest level of granularity obtainable. In cases where such access is needed, it should be provided through a protected and monitored data environment, such as a virtual (remote-access) data enclave,<sup>400</sup> and complemented with data use agreements providing information accountability and appropriate restrictions on use and sharing of the data.

It is clear that OSHA should consider implementing some of these privacy controls when they would provide better privacy and better utility than traditional de-identification techniques like redaction. At the same time, in many cases, having only a single data-sharing model will not suffice for all uses, and thus a tiered access framework can be valuable, and is strictly necessary where a data provider chooses to enable all possible data analyses. Although no form of sharing is completely free of privacy risks, tiered access can be used to provide stronger privacy protections and better utility for different types of uses.<sup>401</sup> The implementation of such a system requires thoughtful analysis with expert consultation to evaluate the uses, threats, and vulnerabilities and to design useful and safe release mechanisms. In addition, a toolkit or other educational materials could offer helpful guidance to assist employers in identifying information within their workplace injury records that poses a disclosure risk, especially if OSHA continues to elicit textual descriptions of injuries and illnesses in the future. Such materials could help reduce the likelihood that employers will include identifying information in the forms they submit to OSHA.

---

399. See, e.g., Dwork et al., *supra* note 314.

400. See Julia Lane & Stephanie Shipp, *Using a Remote Access Data Enclave for Data Dissemination*, 2 INT'L J. DIGITAL CURATION 128 (2007).

401. See National Research Council reports cited *supra* note 17.

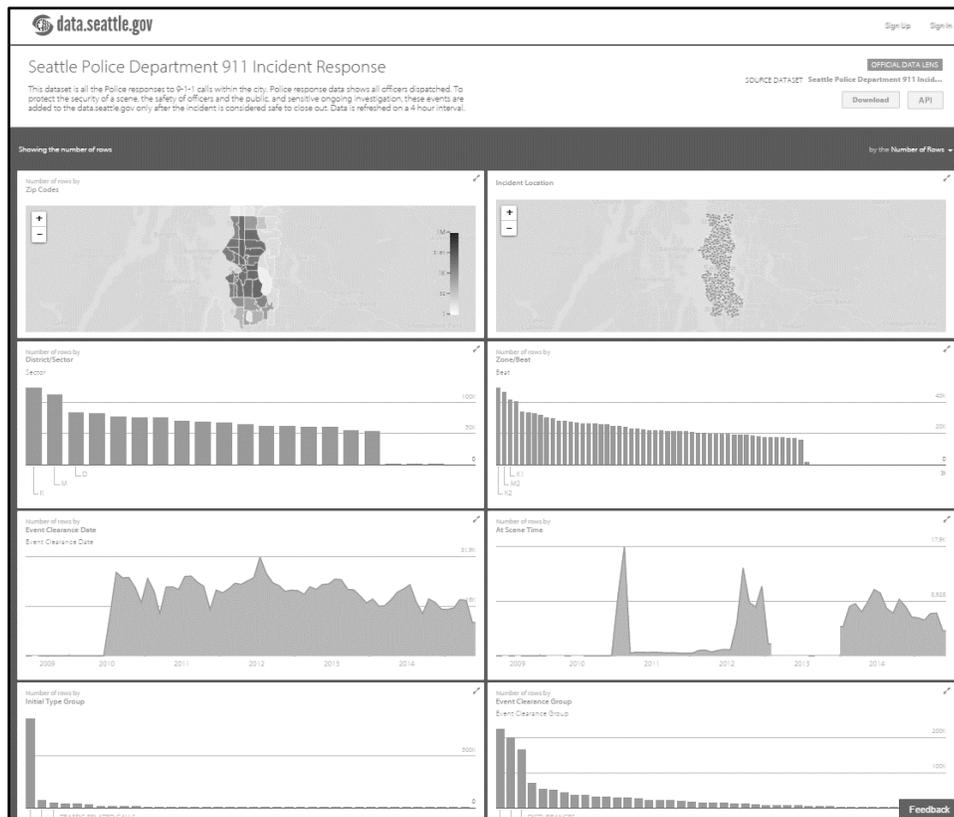


Figure 4: Screenshot of display of the “data lens” visualization of police department 911 incident response data from the City of Seattle open data portal.

## B. MUNICIPAL OPEN DATA PORTALS

Boston and Seattle are two cities that have been rapidly releasing data to the public via open data portals. Through the Socrata open data repository platform, the City of Boston has published over 350 datasets,<sup>402</sup> and the City of Seattle has released over 300 datasets.<sup>403</sup> The cities make their open data available as raw data files, “data lens” interactive visualizations that simplify the interpretation of the raw data (Figure 4),<sup>404</sup> customizable maps and charts, and feeds to an API that enables apps to

402. See *Data Boston*, CITY OF BOSTON, <https://data.cityofboston.gov> (last visited July 15, 2015).

403. See *Results Matching Type of Datasets*, DATA.SEATTLE.GOV, <https://data.seattle.gov/browse?limitTo=datasets> (last visited July 15, 2015).

404. See, e.g., *Seattle Police Department 911 Incident Response*, DATA.SEATTLE.GOV, <https://data.seattle.gov/view/mzrk-e8qt> (last visited May 26, 2015).

query the system and receive real-time streams of data.<sup>405</sup> As discussed below using information learned through interviews with the cities' open data managers, the release of data through these open data portals could be enhanced by systematically aligning intended uses, threats, and vulnerabilities with available privacy controls.

### 1. *Collection and Acceptance Stage*

When initiating data collection, governments should explicitly state the intended uses of the data. The Boston and Seattle open data portals contain data that originated as administrative, statistical, and other records collected for purposes other than release through an open data portal. Examples of these types of records include those related to restaurant licenses, building permits, building and property code violations, census data, constituent services requests, crime reports, 911 emergency calls, and business and professional licenses, which are used by cities to administer agency programs and provide services to city residents.<sup>406</sup> Cities also collect data from other sources, including infrastructure such as utility poles, traffic lights, and streetlamps, which are increasingly being fitted with networked sensors and cameras to collect temperature, light, noise, movement, and emissions data.<sup>407</sup> These data are collected from residents for use by the city, and it is not clear that the subjects of the data are given notice, at the collection stage, that their data will be made available through an open data portal. These are indications that a privacy control at the collection stage, such as transparency about the scope of data collection, use, and release, may be appropriate.

The threats and vulnerabilities associated with the data when collected vary according to the type of information being collected. For instance, information related to business licenses may be identifiable but not very sensitive, while data the cities collect on 911 emergency calls and 311 constituent services calls contain fine-grained information, including date, time, location, and details about the incident and the caller, which may often be both identifiable and sensitive. The latter category of information

---

405. See, e.g., *Seattle Real Time Fire 911 Calls: Official Data Lens*, DATA.SEATTLE.GOV, <https://data.seattle.gov/view/upug-ckch> (last visited May 26, 2015).

406. See *Data Boston: Results matching type of Datasets*, CITY OF BOSTON, <https://data.cityofboston.gov/browse?limitTo=datasets> (last visited Aug. 17, 2015); *Search & Browse Datasets and Views*, DATA.SEATTLE.GOV, <https://data.seattle.gov/browse> (last visited Aug. 17, 2015).

407. See, e.g., STREET BUMP, <http://www.streetbump.org> (last visited Aug. 17, 2015) (describing an app produced by the Mayor's Office of New Urban Mechanics in Boston that uses a smartphone's built-in sensors to detect potholes that volunteers encounter while driving throughout the city).

may be particularly vulnerable to reidentification or inferential risks once it has been collected by the managers of the open data portal. Threats at this stage may include the inadvertent leakage of information by city employees as they collect and process the records. To determine whether additional privacy controls should be implemented at the collection stage, the cities should consider whether the expected benefits of the collection outweigh the potential harms, and whether the broad scope of intended uses would make implementation of certain controls, such as the operationalization of a collection limitation principle, inappropriate. In addition, the city data managers should be transparent to the public about the scope of data collection, the implementation of privacy controls, and the rationales supporting these choices.

### *2. Retention Stage*

The data collected by cities are retained in the information systems of various city departments, and, once they are transferred to the managers of the open data programs, they are stored in a central database. The retention of these records electronically within a central database changes the information security vulnerability surface, and increases the potential for confidentiality loss due to security breach, as a single breach can then compromise a vast quantity of data. Information security controls such as encryption and federated databases are examples of privacy controls that can be implemented to mitigate disclosure risks at the retention stage.

### *3. Post-Retention Transformation*

Open data managers for Boston and Seattle often receive either unaltered data or data that have been redacted or aggregated by the city departments that created the records. In either case, the open data managers review each dataset prior to release to determine whether it contains sensitive information and whether additional aggregation or suppression is needed to mitigate disclosure risks. During this disclosure limitation review, certain identifying fields, such as names, Social Security numbers, and telephone numbers, are typically removed from the data. For example, the City of Seattle removes the address field from business license records before they are published to the open data portal because some businesses are licensed under an owner's home address.<sup>408</sup> At this stage, the open data managers also remove or mask categories of sensitive information. For instance, the City of Boston removes all domestic

---

408. See Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.

violence and sexual assault cases from its crime incident data and generalizes descriptions for the remaining incidents using broad categories such as “drug charges.”<sup>409</sup> In some datasets, incident or call location is coarsened to the block, neighborhood, or city level, and the appropriate granularity is typically chosen by an open data manager on a case-by-case basis. For example, the City of Boston determines whether to generalize location to the block, neighborhood, or city-wide level by transforming the data, viewing the output at each level, and choosing a setting that seems to maximize both utility and privacy.<sup>410</sup> Open data managers also aim to generate metadata that specify which fields were suppressed in a given set of data and the reason for their removal, but it is not always the case that such metadata are created and released with the data.

Some vulnerabilities related to the sensitivity and identifiability of information persist despite efforts to screen, redact, and coarsen the data before release. A set of data that have been generalized and stripped of more specific details may still contain sensitive information. For example, the City of Seattle’s records for 911 incidents include details for events that would generally be considered to be sensitive, such as those categorized as mental illness complaints, drug violations, drug overdoses, prostitution, and lewd behavior.<sup>411</sup> If these general incident descriptors were matched to an identifiable individual using personal knowledge, a news report, or other auxiliary information, it may cause harm to that individual even in the absence of additional details about the incident. Sensitive attributes that are not required to be removed by statute may not be identified as sensitive, or those that appear in only a small subset of the records may be overlooked when reviewing and redacting a dataset before release. Records in the City of Boston’s 311 constituent services requests data include some that are coded as “Breathe Easy” inspections.<sup>412</sup> Breathe Easy at Home is a housing inspection program the city offers for residents who suspect “substandard housing conditions may be triggering a child’s

---

409. See *Crime Incident Reports*, CITYOFBOSTON.GOV, <https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx> (last visited May 26, 2015).

410. See Off-the-record interview with an open data manager for the City of Boston, Apr. 9, 2015.

411. See *Seattle Police Department 911 Incident Response*, DATA.SEATTLE.GOV, <https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp> (last visited Aug. 17, 2015).

412. See *Mayor’s 24 Hour Hotline, Service Requests*, CITY OF BOSTON, <https://data.cityofboston.gov/City-Services/Mayor-s-24-Hour-Hotline-Service-Requests/awu8-dc52> (last visited May 26, 2015).

asthma in their home.”<sup>413</sup> Thus, records associated with this code may reveal that a member of a particular household suffers from asthma. The presence of this sensitive information after transformation is an indication that additional privacy controls, such as more systematic risk assessments and generation of contingency tables, should be explored to better address disclosure risks at the transformation stage.

#### 4. *Release and Access Stage*

Open data managers should also consider the intended uses and expected benefits of open data at the release stage. Information is released to the public as open data to enhance government transparency and accountability and foster greater civic engagement. Such release programs explicitly aim to maximize the quantity of data made available in open formats in order to enable members of the public to find novel and unforeseen uses of the data that will provide benefits for society and promote economic growth. Uses of open data released by Boston and Seattle have included, for example, third party smartphone apps for tracking data points such as 911 calls to local police and fire departments,<sup>414</sup> and feeds for data-driven services like the real estate search engine Zillow. In other words, the intended uses of open data are broadly defined, and the expected benefits of these releases include improvements in service delivery by the government, accountability and transparency for government activities, economic growth, and advances in scientific research. The benefits are intended to accrue to government agencies, commercial entities, researchers, and the public as a whole. For these reasons, the cities should carefully choose privacy controls that support a broad range of analyses and do not unnecessarily preclude uses for which the expected benefits outweigh the privacy risks.

To identify privacy vulnerabilities in a data release, the cities’ open data managers should examine the scope of the information covered. Both Boston and Seattle review, transform, and withhold or release records with the goal of releasing as open data information associated only with minimal privacy risks. For the City of Boston, the scope of release is determined based on guidance from open data policies, such as the mayor’s executive order on open data,<sup>415</sup> and is limited by regulations

---

413. See *Breathe Easy at Home*, CITY OF BOSTON, <http://www.cityofboston.gov/isd/housing/bmc.asp> (last visited May 26, 2015).

414. See DATA.SEATTLE.GOV, <https://data.seattle.gov> (last visited May 26, 2015).

415. Executive Order, Martin J. Walsh, Mayor of Boston, An Order Relative to Open Data and Protected Data Sharing (Apr. 7, 2014), <http://www.cityofboston.gov/news/Default.aspx?id=6589>.

protecting certain categories of information, such as FERPA<sup>416</sup> for education records and state regulations for criminal record information.<sup>417</sup> These are examples of categories of records that the city has a clear duty to protect because the records are expressly protected by law. The scope of information released by the City of Seattle is determined by the State of Washington freedom of information law,<sup>418</sup> which is quite expansive, requiring the release of almost all government records upon request and drawing very narrow exceptions for privacy-sensitive information. Seattle's open data program is also guided by an evolving three-level data classification scheme, describing public data that can be made available without restriction, restricted data that can be released once it has been sanitized, and confidential data that cannot be released to the public at all due to concerns about privacy.<sup>419</sup> Beyond the categories of information the cities have a clear duty to protect, the open data managers express uncertainty regarding how to determine which records should be withheld or redacted as a good practice. Because the cities' open data policies rely on broadly permissive and discretionary state freedom of information laws that prohibit the release of information in only a few narrowly-drawn categories, the cities should consider implementing additional privacy controls at the point of release, such as risk assessments, purpose specification, and transparency, to limit or provide notice of the scope of information released in a systematic way.

Cities should also explicitly identify vulnerabilities arising from the likelihood of reidentification and the learning potential of the data. Current practices for screening data for privacy risks are ad hoc, with open data managers claiming to rely in part on common sense and good judgment to determine whether a given set of data is safe to release through an open data portal. For example, when the open data managers for the City of Seattle receive a dataset from the city department that created the records, they review the columns in the dataset and make a decision as to whether any of the fields likely contain personally identifiable information.<sup>420</sup> They look to regulatory classifications of personally identifiable information from laws such as the HIPAA Privacy Rule; however, they note that these laws are limited in scope and the lack of more comprehensive, formal guidance creates uncertainty. To address

---

416. FERPA, 20 U.S.C. § 1232g (2012); 34 C.F.R. pt. 99 (2015).

417. MASS. GEN. L. ch. 6, § 172 (2015).

418. WASH. REV. CODE § 42.56.001–42.56.904 (2006).

419. *See* Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.

420. *See id.*

these concerns, the City of Seattle is developing new governance procedures, requirements for reviewing and addressing disclosure risks in open data, and definitions for concepts like personally identifiable information.<sup>421</sup>

Data made available through the Boston and Seattle open data portals sometimes contain identifying information. In some cases, a city may have a policy in place for scrubbing data of certain types of identifying information, but, in practice, some fields are overlooked. For example, the City of Boston's 311 constituent request call records contain directly identifying information, such as full street addresses for all calls, and, seemingly inadvertently, include names and telephone numbers for some residents in a field containing free-form text.<sup>422</sup> For some of the 311 records, a field describing the reason for closing a case provides contact information for the caller and contextual details about complaints, which can involve issues related to evictions and homelessness, medical conditions and disabilities, stalking incidents, and interpersonal relationship issues.<sup>423</sup> Some of the fields contain what seem to be lengthy emails from constituents describing their personal situations in great detail and including their own names, addresses, and telephone numbers.<sup>424</sup> For instance, one record describes a domestic dispute involving child custody and visitation violations, restraining orders, and a personal relationship with a registered sex offender, as well as the phone number of the person who called the hotline.<sup>425</sup>

In other cases, a city intends to apply a privacy control but fails to implement it properly,<sup>426</sup> or applies a standard for privacy protection that might not be sufficiently protective for all records. The City of Seattle publishes fire department 911 dispatch data that include a complete address and precise latitude-longitude information for the location of the incident, a coded value for the type of dispatch, and the date and time of

---

421. *See id.*

422. *See Mayor's 24 Hour Hotline Service Requests, supra* note 412.

423. *See Mayor's 24 Hour Hotline, Service Requests, supra* note 412.

424. *See id.*

425. *See id.*

426. For example, several municipal open data portals generalize address fields for crime incident reports to the block level, but also include precise latitude-longitude coordinates that reveal the actual location. *See, e.g.,* Anchorage, Alaska, data at Regional Analysis and Data Sharing (RAIDS) (last visited Aug. 10, 2015), <http://www.raidsonline.com>.

the call.<sup>427</sup> The City of Seattle also publishes police department 911 incident data that include the time the officer arrived on scene, the time the event was cleared, a coded value for the event description, and an address coarsened to the block level.<sup>428</sup> Although police incident data are provided at the block level, if the date, time, and coarsened location are linked with auxiliary information such as that found in a newspaper report, public records database, or social media post, it is likely one could associate the details of some of the incidents with the individuals involved.<sup>429</sup> In addition, a record may be particularly vulnerable to reidentification if it is generalized to a block or other geographic area with a low population density. The presence of potentially identifiable information in the open data portals, despite laws barring the release of certain categories of personal information and stated policies broadly prohibiting the release of identifiable information, is evidence that the programs should seek to screen the data more systematically before release and select appropriate privacy controls at the release stage.

Cities should also consider the threats associated with a data release, which can vary for different types of datasets and different records within datasets that are made available through municipal open data portals. Because the data are open and accessible by anyone, the barrier is low for a neighbor, for instance, to visit an open data portal to learn more about 311 complaints filed by their neighbors or to investigate a recent neighborhood incident to which the police or fire department responded. More sophisticated adversaries, such as data brokers, could mine the data provided through online portals to make inferences about individuals and incidents throughout the city, and these inferences could be used to discriminate against certain populations.<sup>430</sup> As mentioned above, the City of Boston releases the full street addresses of residences that apparently participate in a program to assist individuals suffering from asthma, and a

---

427. See *Seattle Real Time Fire 911 Calls*, DATA.SEATTLE.GOV, <https://data.seattle.gov/Public-Safety/Seattle-Real-Time-Fire-911-Calls/kzjm-xkqj> (last visited May 28, 2015).

428. See *Seattle Police Department 911 Incident Response*, DATA.SEATTLE.GOV, <https://data.seattle.gov/Public-Safety/Seattle-Police-Department-911-Incident-Response/3k2p-39jp> (last visited May 28, 2015).

429. For a demonstration of this type of record linkage, see, e.g., Sweeney, *supra* note 20.

430. See generally Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014) (discussing the discriminatory impact of practices by data brokers and other businesses for mining data from various sources and using predictive algorithms to make credit, employment, insurance, and other decisions).

simple online public records search will likely reveal the names of the individuals residing at those addresses. One might also be able to infer sensitive details such as the socioeconomic status for individuals living at addresses for which complaints of “unsatisfactory living conditions,” “illegal occupancy,” and “overcrowding” have been filed.<sup>431</sup> These are just some examples of the types of threats cities should take into account when designing their data releases and determining which uses they intend to support or prevent. The suitability of privacy controls such as systematic risk assessments, privacy-aware contingency tables and interactive mechanisms, and secure data enclaves, should be explored to reduce the risk that identifiable or sensitive information will be leaked in a municipal open data release.

##### 5. *Post-Access Stage*

When designing an open data release, managers should also consider the threats and vulnerabilities at the post-access stage and select privacy interventions that can address disclosure risks after the data have been released. As noted above, information from an open data portal may be identifiable and sensitive and could be used by a neighbor, friend, family member, potential employer or insurer, or data broker in ways that may cause harm to the subjects of the data. However, once information is published to the Boston and Seattle open data portals, the cities take no further steps to monitor for or prevent misuses of the data and provide no redress for individuals harmed by misuses of the data. While the software used to host the open data portals enables some tracking and monitoring of user actions related to accessing and exploring datasets, the city open data managers have not implemented tools for detecting possible cases of improper use of the data. The open data portals also do not require data users to register, nor do they record an individual’s contact information or attempt to verify one’s identity. The open data managers are therefore unable to contact users who may have accessed data that they should return or destroy because disclosure risks were later discovered. Although the portals provide terms of service that disclaim responsibility in areas such as data accuracy, they do not specify restrictions on use, expressly prohibit users from attempting to reidentify individuals in the data, require users to notify city data managers of disclosure risks discovered in the data, or specify enforcement or accountability mechanisms for misuse of the data. These types of provisions are among the most common restrictions and requirements found in the terms of use for other large data

---

431. *See id.*

repositories, and ones the cities should consider incorporating into their policies in order to mitigate disclosure risks at the post-access stage.<sup>432</sup>

6. *Aligning Use, Threats, and Vulnerabilities with Controls*

As discussed above, the Boston and Seattle open data portals rely on withholding, redacting, and, to a lesser extent, coarsening information deemed to be sensitive or identifying before release. The procedures they use to review data for privacy risks are ad hoc and typically involve one or two data managers reviewing the columns of a dataset for obvious direct and indirect identifiers.<sup>433</sup> Likewise, in transforming the data they rely on heuristics rather than formal standards to redact fields or collapse values into large categories.<sup>434</sup> In a few cases, the cities release data received from city departments as summary files. For example, City of Boston census data are released as contingency tables describing demographic characteristics of various city neighborhoods, rather than raw, individual-level data.<sup>435</sup> Visualization tools are often provided to make data analysis simpler and more intuitive for visitors to the web-based portal, but such tools do not incorporate privacy-preserving features such as aggregation and noise addition. The City of Boston, for example, provides a tool for mapping 311 calls across the city, and, although it aggregates information in the displayed map, this is done for ease of analysis rather than for privacy, as it also includes all of the raw, individual-level data in a table displayed below the map (Figure 5). We could not find any examples of the open data portals making use of more advanced techniques for privacy protection, such as privacy-aware contingency tables, visualizations, or interactive mechanisms.

---

432. For an example of standard terms of use implemented by one of the largest data repositories, see, for example, The Interuniversity Consortium for Political and Social Research, *What Are ICPSR's Terms of Use?* (2009), <http://www.icpsr.umich.edu/icpsrweb/membership/support/faqs/2009/01/what-are-icpsrs-terms-of-use>.

433. See Off-the-record interview with open data managers for the City of Seattle, May 21, 2015; Off-the-record interview with an open data manager for the City of Boston, Apr. 9, 2015.

434. See sources cited *supra* note 433.

435. See, e.g., *South Boston, neighborhood: 2010 Census*, CITY OF BOSTON, <https://data.cityofboston.gov/dataset/South-Boston-neighborhood-2010-Census/ybpb-72n5> (last visited May 26, 2015).

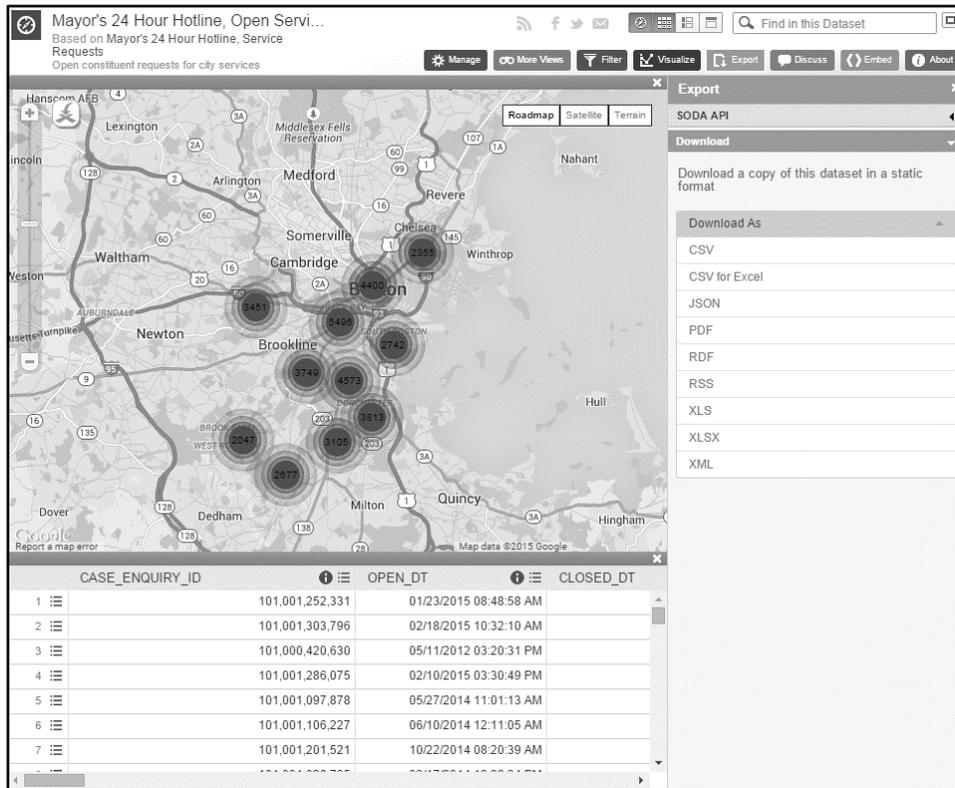


Figure 5: Visualization of 311 data in the City of Boston open data portal.

The privacy threats and vulnerabilities in the Boston and Seattle open data portals and the lack of formal standards and procedures for screening data and employing privacy controls point to the need for a more systematic approach to assessing privacy risks and implementing appropriate privacy controls in these programs. In fact, efforts are currently underway to move in that direction. The City of Boston is developing a decision tree to assist departments in classifying their data using a more systematic risk assessment model, as well as complementary data policy guidance for agencies and departments across the city to the follow.<sup>436</sup> Similarly, the City of Seattle's data managers are currently engaged in a process to develop more rigorous governance procedures for its open data program and to draft new rules and policies for classifying, de-identifying, and releasing data.<sup>437</sup> In these efforts, the cities should

436. See Off-the-record interview with an open data manager for the City of Boston, Apr. 9, 2015.

437. See Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.

consult with data privacy experts to ensure that the new standards and procedures take into account recent advances in privacy from fields such as computer science, statistics, and law.

In particular, the cities' open data portals would likely see gains in both privacy and utility with the adoption of a tiered access model for data containing sensitive information. Tiered access, as described more fully in the OSHA case in Section IV.A, allows for the implementation of privacy controls that are finely tuned to the intended uses, threats, and vulnerabilities relevant to the release. In this way, agencies can support a variety of uses of the data while providing robust privacy protections for the individuals in the data. Such a model would seem to be particularly well-suited for open data portals, which are intended to support a broad range of uses across different types of data. For example, an open data portal could enable public access to privacy-preserving contingency tables and visualizations for certain types of data such as 911 calls, for which accuracy at the neighborhood level may be adequate to serve the needs of users who are tracking public safety. At an intermediate level of access, cities could make data available through an interactive mechanism, which could also enable analysis of information that is currently stripped from data, including finer grained location information or certain types of sensitive records such as sexual assault incidents. When researchers need full access to the data, the data could be made available to approved users through a virtual data enclave, under the terms of a data use agreement. In addition to these controls at the release stage, the cities should also consider adopting controls on the collection and storage of information, as well as post-release review, accountability, and redress mechanisms to monitor and detect misuses of data and enable enforcement in response to privacy breaches.

## V. SUMMARY

There is a growing consensus among privacy scholars, policymakers, and the public that common approaches to privacy are incomplete and inconsistent. In response many approaches and interventions have been proposed, but the result is a landscape of privacy regulation and policy recommendations that has become increasingly difficult to navigate and understand.

In this Article, we examine the area of privacy regulation for government data releases and plot a path through its terrain. We analyze how information is currently treated from cradle to grave within major categories of government releases of data, and contrast that treatment with the wide range of considerations and interventions suggested in scholarly

analyses of privacy. What we find from an examination of broad categories of release mechanisms and specific data release cases both reinforces current concerns and outlines a framework for approaching regulatory solutions.

For instance, we find that the treatment of privacy across different types of data releases is highly inconsistent. In some cases, identical information, measuring the same characteristics of the same people, is subject to very different assessments of privacy risk and selections of privacy controls, merely because the information is being distributed through different endpoints. Moreover, and, more commonly, sets of data that pose the same risks to the same types of data subjects are treated vastly differently. In other words, the criteria considered most relevant to privacy protection by the scholarly and policy community appear to be generally absent from regulations and practices on the ground. In addition, there is very little guidance available to agencies regarding the application of regulatory standards for privacy protection in specific circumstances, and this contributes to inconsistencies in practice and potential ineffectiveness of privacy safeguards adopted.

We find also that there are many gaps in the privacy controls used in government data release programs. The scholarly and policy literature has identified a wide range of technical, procedural, legal, educational, and economic controls; however, for the most part, government data releases rely entirely on redaction and binary access control. This focus on a small set of controls likely fails to address the nuances of data privacy risks. It also stands in contrast to the practice of information security, which involves the implementation of a wide range of security controls from a diverse, organized, and well-documented catalog.

Addressing privacy risks requires a sophisticated approach, and the privacy protections currently used in government releases of data do not take advantage of advances in data privacy research or the nuances these provide in dealing with different kinds of data and closely matching privacy controls to the intended uses, threats, and vulnerabilities of a release. Combined with a review of the broader literature and existing high-level principles for privacy protection, we propose a framework for developing appropriate data release mechanisms for particular cases such as the public release of OSHA-collected workplace injury records and the release of records through municipal open data portals. By tracing the information involved in government data releases, we identify five distinct operational stages: collection, transformation, retention, transformation, release, and post-access. At each of these stages, there are a number of factors related to the intended uses, threats, and vulnerabilities that should

be considered when developing an appropriate data release mechanism. In addition, at each stage policymakers have the opportunity to select from a distinct set of legal, technical, economic, procedural and educational interventions, in order to construct a comprehensive policy. The selection of controls should be calibrated to the specific uses, threats, and vulnerabilities identified.

In the rapidly changing environment of information policy and technology, neither science nor principle provides definitive guidance on how to select policy components for a data release based on the risks and benefits of each case. At the same time, changes in science and technology offer the opportunity for sophisticated characterization of privacy risks and harms, and more modern forms of educational interventions and technical controls. An information lifecycle framework, while not yet fully prescriptive, can provide a systematic and useful decomposition of the factors relevant to data release, and can be used to order the set of interventions that should be considered at each stage. Further, a systematic framework provides a natural foundation for increased transparency, and we encourage government actors to be transparent in documenting the uses, potential risks, and the privacy and security interventions selected at each lifecycle stage.

# OPEN DATA, PRIVACY, AND FAIR INFORMATION PRINCIPLES: TOWARDS A BALANCING FRAMEWORK

*Frederik Zuiderveen Borgesius, Jonathan Gray & Mireille van Eechoud<sup>†</sup>*

## ABSTRACT

Open data are held to contribute to a wide variety of social and political goals, including strengthening transparency, public participation and democratic accountability, promoting economic growth and innovation, and enabling greater public sector efficiency and cost savings. However, releasing government data that contain personal information

---

DOI: <http://dx.doi.org/10.15779/Z389S18>

© 2015 Frederik Zuiderveen Borgesius, Jonathan Gray & Mireille van Eechoud.

<sup>†</sup> Frederik Zuiderveen Borgesius is a post-doctoral researcher at the Institute for Information Law, University of Amsterdam Law School, The Netherlands. Jonathan Gray is a Research Associate at the Digital Methods Initiative, University of Amsterdam and Director of Policy and Research at Open Knowledge. Mireille van Eechoud is professor of Information Law at the Institute for Information Law, University of Amsterdam Law School.

We thank Simon Hania, Dariusz Kloza, Stefan Kulk, Maja Lubarda, Richard Rogers, Javier Ruiz, Nico van Eijk, Ben Worthy, and Bendert Zevenbergen for participating in the *Workshop Reconciling Fair Information Principles and Open Data Policies* on February 6, 2015 at the Institute for Information Law, Amsterdam. We also thank the participants of the symposium *Open Data: Addressing Privacy, Security, and Civil Rights Challenges*, April 17, 2015, Berkeley Center for Law & Technology, in particular Cathy O'Neil and David Flaherty. The thought-provoking discussions during both events helped to shape our ideas for this Article. Furthermore, Matthijs Koot, Bendert Zevenbergen, and the editors of the Berkeley Technology Law Journal deserve our gratitude for comments on earlier versions of this Article. We also express our gratitude to the members of the advisory board for the project that led to this Article: Simon Hania, Corporate Privacy Officer at the TomTom company; Dr. Jaap-Henk Hoepman, Associate Professor of Privacy Enhancing Protocols and Privacy by Design, University of Nijmegen; Dr. Aleecia McDonald, non-residential fellow, Center for Internet & Society, Stanford University; Prof. B. Roessler, Professor of Ethics and its History, University of Amsterdam; Javier Ruiz Diaz, Policy Director, Open Rights Group; Prof. N.A.N.M. van Eijk, Professor of Media and Telecommunications Law, University of Amsterdam; Dr. Ben Worthy, lecturer in Politics at Birkbeck University of London, independent reporter for the U.K.'s IRM of the Open Government Partnership (U.K.). We thank Sarah Eskens, Rachel Wouda, and Dirk Henderickx for research assistance. All errors are the authors' own. Financial support for this project came from the Berkeley Center for Law & Technology and Microsoft.

may threaten privacy and related rights and interests. In this Article we ask how these privacy interests can be respected, without unduly hampering benefits from disclosing public sector information. We propose a balancing framework to help public authorities address this question in different contexts. The framework takes into account different levels of privacy risks for different types of data. It also separates decisions about access and re-use, and highlights a range of different disclosure routes. A circumstance catalogue lists factors that might be considered when assessing whether, under which conditions, and how a dataset can be released. While open data remains an important route for the publication of government information, we conclude that it is not the only route, and there must be clear and robust public interest arguments in order to justify the disclosure of personal information as open data.

## TABLE OF CONTENTS

I.	INTRODUCTION .....	2075
II.	OPEN DATA AND PRIVACY .....	2078
A.	OPEN DATA INTERESTS .....	2078
1.	<i>Innovation and Economic Growth</i> .....	2080
2.	<i>Political Accountability and Democratic Participation</i> .....	2083
3.	<i>Public Sector Efficiency and Service Delivery</i> .....	2085
B.	PRIVACY INTERESTS .....	2086
1.	<i>Chilling Effects</i> .....	2088
2.	<i>Lack of Control over Personal Information</i> .....	2089
3.	<i>Social Sorting and Discrimination</i> .....	2091
III.	GOVERNANCE OF PUBLIC SECTOR INFORMATION .....	2093
A.	OPEN DATA NORMS .....	2093
B.	ACCESS TO INFORMATION NORMS .....	2095
C.	ACCESS TO INFORMATION NORMS AND PRIVACY .....	2098
IV.	GOVERNANCE OF PERSONAL INFORMATION .....	2101
A.	FAIR INFORMATION PRINCIPLES (FIPS) .....	2101
1.	<i>Background of the FIPs</i> .....	2101
2.	<i>OECD Guidelines</i> .....	2102
3.	<i>Scope of the OECD Guidelines</i> .....	2106
B.	FIPS AND OPEN DATA: CHALLENGES .....	2107
1.	<i>Purpose Specification Principle</i> .....	2109
2.	<i>Security and Accountability Principles</i> .....	2110
3.	<i>Data Quality Principle</i> .....	2111
4.	<i>Collection Limitation and Transparency Principle</i> .....	2111
5.	<i>Use Limitation and Individual Participation Principle</i> .....	2112
V.	TYPES OF DATA .....	2114
A.	RAW PERSONAL DATA .....	2114
B.	PSEUDONYMIZED DATA .....	2116
C.	ANONYMIZED DATA .....	2118

D.	NON-PERSONAL DATA .....	2120
E.	FUZZY BOUNDARIES .....	2121
VI.	TYPES OF DISCLOSURE .....	2122
A.	DISCLOSURE WITH ACCESS RESTRICTIONS .....	2122
B.	DISCLOSURE WITH RE-USE RESTRICTIONS .....	2124
C.	DISCLOSURE AS OPEN DATA.....	2125
VII.	A CIRCUMSTANCE CATALOGUE TO INFORM DISCLOSURE DECISIONS.....	2125
A.	WEIGHT OF THE GOALS PURSUED.....	2126
B.	WEIGHT OF THE PRIVACY INTERESTS .....	2128
VIII.	CONCLUSION.....	2129

## I. INTRODUCTION

Open government data refers to data released by public sector bodies, in a manner that is legally and technically re-usable. The *G8 Open Data Charter* states “free access to, and subsequent re-use of, open data are of significant value to society and the economy.”<sup>1</sup> Open data are commonly held by its advocates to mean data that “can be freely used, modified, and shared by anyone for any purpose.”<sup>2</sup> However, releasing public sector datasets that include personal information, or data that can be re-identified, may threaten privacy and related rights.

In this Article, we examine the tension between public sector open data policy and the Fair Information Principles (FIPs). The FIPs lie at the core of most data privacy laws around the world, including those in the European Union and the United States. The FIPs give guidelines to balance privacy-related interests and other interests, such as those of business and the public sector. The Article focuses on the following question: from the perspective of the Fair Information Principles, how can privacy and related interests be respected, without unduly hampering benefits from disclosing public sector information?

We rely mostly on desk research, using the usual sources for legal scholarship, such as legislation, soft law, policy documents, and literature.

---

1. G8 OPEN DATA CHARTER (2013) <https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex>.

2. OPEN DEFINITION, <http://opendefinition.org> (last visited May 1, 2015). Open Knowledge’s first open definition dates from 2005. *Open Knowledge Definition 1.0*, OPEN DEFINITION, <http://opendefinition.org/od/1.0> (last visited May 1, 2015).

We use descriptive and analytical legal research to determine the main legal tensions between open data policy and the FIPs. Parts of the Article are more normative: we give recommendations to strike a balance that respects privacy and related interests, and does not unduly hamper the benefits of open data.

We enriched our research results with insights from a workshop, where we tested hypotheses and discussed the promises and pitfalls of privacy and open data. Conference participants came from academia, industry, civil society organizations, and data protection authorities, and were all working on issues in open data and privacy.<sup>3</sup> Discussions during the *Open Data: Addressing Privacy, Security, and Civil Rights Challenges* symposium held by the Berkeley Center for Law & Technology also provided valuable insights.<sup>4</sup>

Furthermore, we conducted an empirical study into concerns that various stakeholders, in civil society, the public sector, research, and business, express about the interactions between privacy and open data. The study draws on document collections and digital traces from the web to map the debates about privacy and open data. The empirical study follows the “digital methods” approach, pioneered by Richard Rogers and his colleagues at the Digital Methods Initiative.<sup>5</sup>

While each national legal system has its own traditions and characteristics, this Article focuses on common problems that arise in many jurisdictions. After all, as the Open Government Partnership (OGP) testified, governments around the world create open data policies and must cope with privacy concerns.<sup>6</sup> Hence, we do not examine what sets jurisdictions apart, but instead discuss shared problems. For instance, we do not address specific requirements that follow from the First

---

3. Workshop, Reconciling Fair Information Principles and Open Data Policies, Institute for Information Law, Amsterdam, Netherlands (Feb. 6, 2015).

4. See *Addressing Privacy, Security, and Civil Rights Challenges—19th Annual BCLT/BTLJ Symposium*, BERKELEY LAW (Apr. 17, 2015), <https://www.law.berkeley.edu/centers/bclt/past-events/april-2015-the-19th-annual-bcltbtlj-symposium-open-data-addressing-privacy-security-and-civil-rights-challenges/program>.

5. See generally RICHARD ROGERS, DIGITAL METHODS (2013).

6. The OGP is an international platform for reform, to make “governments more open, accountable, and responsive to citizens.” Participating states submit action plans in which they make commitments, inter alia on datasets to be made available as open data. Compliance and progress mechanisms are in place. Membership has grown to 65 countries in the five years since the OGP’s inception. See OPEN GOVERNMENT PARTNERSHIP, <http://www.opengovpartnership.org> (last visited May 1, 2015).

Amendment in the United States,<sup>7</sup> or from the fundamental right to data protection in the European Union.<sup>8</sup> Therefore, the Article's recommendations come with a caveat: they cannot be directly implemented in national legal systems.

The Article is structured as follows. Part II describes open data goals and privacy problems regarding open data. We clustered the objectives associated with open data into three categories: (1) innovation and economic growth, (2) political accountability and democratic participation, and (3) public sector efficiency. We identified three kinds of concerns about releasing personal information as open data: (1) the chilling effects on people interacting with the public sector, (2) a lack of individual control over personal information, and (3) the use of open data for social sorting or discriminatory practices.

Part III discusses rules regarding access to information held by public sector. Freedom of information laws provide inspiration on how to strike a balance between privacy and transparency in the open data context.

Part IV discusses the governance of personal information, focusing on the Fair Information Principles (FIPs). In this section we also discuss the main challenges in reconciling open data policy and the FIPs. From a FIPs perspective, the main problem with open data is that unrestricted re-use of personal data breaches the purpose specification principle. But we argue that there are possible compromise measures to balance privacy and open data interests.

We propose a balancing framework to accommodate privacy concerns and open data goals. Part V outlines the first element of the balancing framework, and distinguishes four data categories with different levels of privacy risks: (A) raw personal data, (B) pseudonymized data, (C) anonymized data, and (D) non-personal data. Different modes of access and re-use control are the second element of the balancing framework. In many cases, disclosing data with access or re-use restrictions, rather than as fully open data, strikes a balance between open data goals and privacy (Part VI). As a third element of the balancing framework we provide a circumstance catalogue, a list of circumstances to consider when deciding

---

7. See Daniel J. Solove, *Access and Aggregation: Public Records, Privacy and the Constitution*, 86 MINN. L. REV. 1137, 1201 (2002).

8. For more on EU data protection law and public sector information re-use policy, see Cristina Dos Santos et al., *On Privacy and Personal Data Protection*, 6 MASARYK Ů. J.L. & TECH. 337 (2012), <https://journals.muni.cz/mujlt/article/view/2613/2177>; see also Mireille van Eechoud et al., *LAPSI Position Paper on Access to Data*, LAPSI (Dec. 12, 2014), <http://dare.uva.nl/document/2/162858>.

whether or not a dataset should be disclosed, and under which conditions (Part VII).

Part VIII concludes that releasing personal information as fully open data is generally not appropriate. But sometimes a compromise can be found by disclosing data with access or re-use restrictions.

## II. OPEN DATA AND PRIVACY

Open data are held to contribute to a wide variety of social and political goals. However, releasing data as open data may threaten privacy, for instance, if the open data contain personal information. Below we describe open data goals and privacy problems regarding open data. We clustered the objectives associated with open data into three categories: (1) innovation and economic growth, (2) political accountability and democratic participation, and (3) public sector efficiency. We also clustered privacy concerns in the area of open data into three categories: (1) the chilling effects on people interacting with the public sector, (2) a lack of individual control over personal information, and (3) the use of open data for social sorting or discriminatory practices.

### A. OPEN DATA INTERESTS

Definitions of open data from technologists and civil society actors focus on enabling redistribution and re-use, and on limiting legal and technical barriers to re-use. For example, the summary of the “Open Definition” from Open Knowledge reads: “Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness).”<sup>9</sup> The full definition stipulates conditions that include legal openness, bulk downloadability, and machine-readability.<sup>10</sup> Similar definitions are used in the *8 Principles of Open Government Data*,<sup>11</sup> the Sunlight Foundation’s *Ten Principles for Opening Up Government Information*,<sup>12</sup> and the World Wide Web Consortium’s *Five Stars of Linked Open Data*.<sup>13</sup>

---

9. OPEN DEFINITION, *supra* note 2.

10. *Id.*

11. OPEN DATA WORKING GROUP, *The 8 Principles of Open Government Data*, OPENGOVDATA.ORG (Dec. 8, 2007), <http://opengovdata.org>.

12. *Ten Principles for Opening Up Government Information*, SUNLIGHT FOUNDATION (Aug. 11, 2010), <http://sunlightfoundation.com/policy/documents/ten-open-data-principles>.

13. Tim Berners-Lee, *Linked Data*, W3.ORG (June 18, 2009), <http://www.w3.org/DesignIssues/LinkedData.html>.

Technical obstacles for re-using data include non-machine readable formats, proprietary formats, technological protection mechanisms,<sup>14</sup> and Digital Rights Management software. Legal restrictions on re-use include intellectual property rights, such as copyright and database rights.<sup>15</sup> When open data advocates say that “anyone can freely access, use, modify, and share [data] for any purpose,”<sup>16</sup> they are often referring to removing these specific kinds of legal and technical restrictions.

This conception of open data that focuses on limiting legal and technical restrictions for re-use has carried into public policy. Over the past decade, open data developed from being a niche idea at the margins of open source software, scientific research and hacker communities, into an idea with traction among public policymakers.<sup>17</sup> For example, the 2013 *G8 Open Data Charter* mentions that open data should be “machine readable,” available in bulk, available in formats for which the specification is “available to anyone for free,” and under open licenses such that “no restrictions or charges are placed on the re-use of the information for non-commercial or commercial purposes.”<sup>18</sup> A similar focus on removing technical restrictions to re-use can be found in open data guidelines of the Organisation for Economic Co-operation and Development,<sup>19</sup> the U.K. government,<sup>20</sup> and U.S. President Barack Obama.<sup>21</sup>

---

14. Technological protection mechanisms (TPMs) and digital rights management information are protected against circumvention and interference in their own right, separate from, e.g., copyright in the underlying work (database, software or other works). See Berne Convention for the Protection of Literary and Artistic Works, art. 11, 12, as amended Sept. 28, 1979, S. TREATY DOC. No. 99-27; WIPO Copyright Treaty, Dec. 20, 1996, S. TREATY DOC. No. 105-17.

15. There is controversy about the role of intellectual property rights in implementing public sector open data, but this controversy is beyond the scope of this Article.

16. OPEN DEFINITION, *supra* note 2.

17. Jonathan Gray, *Towards a Genealogy of Open Data* (Sept. 3, 2014) (Conference Paper given at the General Conference of the European Consortium for Political Research in Glasgow, Scotland), <http://dx.doi.org/10.2139/ssrn.2605828>.

18. G8 OPEN DATA CHARTER, *supra* note 1.

19. See Barbara Ubaldi, *Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives* (Organisation for Economic Co-operation & Development, Working Paper on Public Governance No. 22, 2013), <http://dx.doi.org/10.1787/5k46bj4f03s7-en>; see also Organisation for Economic Co-operation & Development [OECD], *OECD Recommendation of the Council for Enhanced Access and More Effective Use of Public Sector Information*, OECD Doc. C(2008)36 (2008), <https://www.oecd.org/sti/44384673.pdf> [hereinafter *OECD Recommendation*].

20. *Public Data Principles*, DATA.GOV.UK (Apr. 10, 2012), <http://data.gov.uk/library/public-data-principles>.

21. See Exec. Order No. 13,642, *Making Open and Machine Readable the New Default for Government Information*, 78 Fed. Reg. 28111 (May 9, 2013), <https://www>

Open data are held to contribute to a wide variety of social and political goals.<sup>22</sup> For ease of discussion in this Article, we have clustered the many objectives associated with open data into the following three areas: (1) innovation and economic growth, (2) political accountability and democratic participation, and (3) public sector efficiency. First we look at fostering innovation and economic growth.

### 1. *Innovation and Economic Growth*

Most official open data initiatives highlight the potential of enabling the re-use of public sector information to create new businesses and innovative services and products. Open data policies are increasingly becoming the preferred route to unlock the value of public sector information. This is evident from the European Commission's Guidelines on the Public Sector Information Directive.<sup>23</sup> President Obama's 2013 executive order, which aims to make Open and Machine Readable the New Default for Government Information, views (federal) government information as a national asset and recognizes the importance of enabling widespread re-use for "economic growth and job creation."<sup>24</sup> President Obama's 2013 executive order on Open Data Policy adds: "making information resources accessible, discoverable, and usable by the public can help fuel entrepreneurship, innovation, and scientific discovery."<sup>25</sup> Similarly, the *G8 Open Data Charter* claims open data are "a catalyst for innovation in the private sector, supporting the creation of new markets, businesses, and jobs."<sup>26</sup> The World Bank also recognizes this potential of open data.<sup>27</sup>

---

.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government- [hereinafter Exec. Order, Open and Machine Readable].

22. See, e.g., Gray, *supra* note 17.

23. Commission Notice: Guidelines on Recommended Standard Licenses, Datasets and Charging for the Re-Use of Documents, 2014 O.J. (C 240) 1.

24. Exec. Order, Open and Machine Readable, *supra* note 21. The Order is one of several that follow up on open government policy announced by the White House in January 2009. Memorandum for the Heads of Executive Departments and Agencies on Transparency and Open Government, 74 Fed. Reg. 15 (Jan. 21, 2009).

25. OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB MEMORANDUM M-13-13, OPEN DATA POLICY—MANAGING INFORMATION AS AN ASSET (2013), <http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf> [hereinafter OMB MEMORANDUM M-13-13, OPEN DATA POLICY].

26. See G8 OPEN DATA CHARTER, *supra* note 1. It was signed by G8 leaders on June 18, 2013 to promote transparency, innovation, and accountability.

27. WORLD BANK, OPEN DATA FOR ECONOMIC GROWTH 5 (2014).

Information services built on public sector data are diverse. Financial services providers use official statistics as input.<sup>28</sup> Companies in the meteorological sector use weather data to provide highly specialized services, e.g., forecasts for off-shore oil industries.<sup>29</sup> Planning permissions, zoning data and housing data are combined with other sources to produce advice for customers such as real estate developers.<sup>30</sup> Postal codes are widely used as identifiers.<sup>31</sup> School and health inspection data serve as input for apps that help inform parents or patient choice.<sup>32</sup> Public transport timetable data when combined with geolocation data enable real-time and customized travel advice.<sup>33</sup> There are many other kinds of commercial exploitation of open data, often involving the combination of data from different public and private sources to deliver information products or services. The emphasis on economic benefits of re-using data held by public sector bodies predates open data policies. For example, in 1989, the E.U. sought to stimulate commercial exploitation of public sector data by the private sector.<sup>34</sup> The E.U. Public Sector Information Directive of 2003 also focused on public sector information as raw material for creating services and products.<sup>35</sup> The Directive obliged a wide range of public sector bodies to allow commercial and non-commercial re-

---

28. For examples of government information re-use, see MARTIN FORNEFELD ET AL., MICUS, REPORT FOR THE EUROPEAN COMMISSION, ASSESSMENT OF THE RE-USE OF PUBLIC SECTOR INFORMATION (PSI) (2009); MAKX DEKKERS ET AL., MEASURING EUROPEAN PUBLIC SECTOR INFORMATION RESOURCES (MEPSIR), REPORT FOR THE EUROPEAN COMMISSION, FINAL REPORT OF STUDY ON EXPLOITATION OF PUBLIC SECTOR INFORMATION 37 (2006).

29. For example, consider the private company MeteoGroup. See *Marine*, METEOGROUP, <http://www.meteogroup.com/en/gb/sectors/marine.html> (last visited May 15, 2015).

30. For example, in Europe, the company Landmark provides such services and took the city of Amsterdam to court for the price it charged for re-use of city data. See ABRvS 20 april 2009, AB 2009, 546 m.nt. JJB (B&W Amsterdam/Landmark) (Neth.).

31. For this reason the G8 Open Data Charter lists postal codes as “high value” data, to be made available with priority. G8 OPEN DATA CHARTER, *supra* note 1.

32. U.S. DEP’T OF COMMERCE, FOSTERING INNOVATION, CREATING JOBS, DRIVING BETTER DECISIONS: THE VALUE OF GOVERNMENT DATA (2014); G8 OPEN DATA CHARTER, *supra* note 1; MCKINSEY & CO., OPEN DATA: UNLOCKING INNOVATION AND PERFORMANCE WITH LIQUID INFORMATION 11 (2013).

33. See MCKINSEY & CO., *supra* note 32, at 6.

34. Directorate Gen. for Telecomm., Info. Indus. & Innovation, Comm’n of the European Cmtys., *Guidelines for Improving the Synergy Between the Public and Private Sectors in the Information Market* (1989).

35. Directive 2003/98/EC, of the European Parliament and of the Council of 17 November 2003 on the Re-use of Public Sector Information, 2003 O.J. (L 345) 90 (revised by Directive 2013/37/EC, 2013 O.J. (L 175) 1).

use of their information assets, but not necessarily as open data.<sup>36</sup> Under the directive, conditions may be imposed, costs charged, and data may be made available in non-structured form. The U.S. Office of Management & Budget also recognized federal information as a “commodity in the marketplace.”<sup>37</sup>

Many studies have been commissioned to assess the value of public sector information; these studies suggest impressive figures, but range widely.<sup>38</sup> For example, the U.S. Department of Commerce looked at the size of private sector revenues from “government data-intensive business activities” for the United States and arrived at a crude estimate in the range of 24 to 221 billion USD per year.<sup>39</sup> And a 2000 study for the European Commission estimated that for the then 15 E.U. member states, the part of the combined national income attributable to industries and activities built on exploiting public sector information ranged between €28 billion and €134 billion. Some have judged these estimates as far too

---

36. The *obligation* to allow re-use was introduced in the 2013 revision. Directive 2013/37/EC, 2013 O.J. (L 175) 1. Member states must implement the revised directive by July 2015. The Directive builds on public access regimes in member states; it does not regulate access directly.

37. Office of Mgmt. & Budget, Exec. Office of the President, OMB Circular No. A-130 Revised, MANAGEMENT OF FEDERAL INFORMATION RESOURCES (1998). First issued in 1985, the Circular fostered (among many things) a larger role for the private sector in dissemination of government information and creating added-value (electronic) services. With subsequent revisions (1993–1996) under the Clinton administration the focus moved to release of electronic information by federal agencies directly to the public. For an overview of early policy development, see U.S. OFFICE OF TECHNOLOGY ASSESSMENT, OTA-C IT-396, INFORMING THE NATION: FEDERAL INFORMATION DISSEMINATION IN AN ELECTRONIC AGE (1988) [*hereinafter* FEDERAL INFORMATION DISSEMINATION IN AN ELECTRONIC AGE].

38. For recent examples of studies on the economic value of public sector information at the E.U. level, see MARC DE VRIES ET AL., PRICING OF PUBLIC SECTOR INFORMATION. MODELS OF SUPPLY AND CHARGING FOR PUBLIC SECTOR INFORMATION, FINAL REPORT (2011); MARC DE VRIES ET AL., REPORT FOR EUROPEAN COMMISSION, PRICING OF PUBLIC SECTOR INFORMATION STUDY (POPSIS) (2011). For recent examples of studies about the value of open data and public sector information at the national level, see U.S. DEP’T OF COMMERCE, *supra* note 32; DELOITTE, MARKET ASSESSMENT OF PUBLIC SECTOR INFORMATION, STUDY FOR U.K. DEPARTMENT FOR BUSINESS, INNOVATION, & SKILLS (2013); U.K. OFFICE OF FAIR TRADING, OFT861, THE COMMERCIAL USE OF PUBLIC INFORMATION (2006). For examples of subnational level studies, see Jens PREISCHE, DIGITALES GOLD: NUTZEN UND WERTSCHÖPFUNG DURCH OPEN DATA FÜR BERLIN (2014); Gregor Eibl & Brigitte Lutz, *Money for Nothing—Data for Free: Hard Facts About the Economic Power of Open Government Data*, in CEDEM13: CONFERENCE FOR E-DEMOCRACY AND OPEN GOVERNMENT 289 (Peter Parycek & Noelle Edelmann eds., 2d ed. 2013).

39. U.S. DEP’T OF COMMERCE, *supra* note 32.

optimistic.<sup>40</sup> Generally, researchers recognize there is a lack of hard data on which to base estimates.<sup>41</sup> Nevertheless, policymakers see fostering innovation and economic growth as an important goal of open data.

## 2. *Political Accountability and Democratic Participation*

A second goal pursued through open data policy is fostering political accountability and democratic participation. Current proactive disclosure policies cover a broad range of information: from basic information about a public authority's responsibility, organization, and procedures, to granular data about public spending and subsidies awarded.<sup>42</sup>

In the open data context, statements about the perceived benefits of open data for democracy are frequent. The *G8 Open Data Charter* mentions good governance and anti-corruption,<sup>43</sup> and argues that more public data on the use of natural resources and distribution of revenues, on land management, and on development spending would promote accountability and good governance.<sup>44</sup> The World Bank makes a similar case, arguing that open data "supports democratic societies" and "encourages greater citizen participation in government affairs."<sup>45</sup> The French government's open data policy is driven by the idea that "opening and sharing data is the way for modern government to organize itself so that it is accountable, opens dialogue and trusts the collective intelligence of its citizens."<sup>46</sup> The Obama administration posits that making information available proactively online in open formats increases

40. Robbin te Velde, *Public Sector Information: Why Bother?*, in *THE SOCIO-ECONOMIC EFFECTS OF PUBLIC SECTOR INFORMATION ON DIGITAL NETWORKS: TOWARD A BETTER UNDERSTANDING OF DIFFERENT ACCESS AND REUSE POLICIES: WORKSHOP SUMMARY 25*, 25–28 (P. Uhler ed., 2009).

41. See Mireille van Eechoud, *Calculating and Monitoring the Benefits of Public Sector Information Re-use*, in *ZUGANG UND VERWERTUNG ÖFFENTLICHER INFORMATIONEN* (Thomas Dreier et al. eds., forthcoming 2015).

42. See, e.g., *Cabinet Office Organogram*, DATA.GOV.UK, <http://data.gov.uk/organogram/cabinet-office>; *Senior Officials "High Earners" Salaries*, DATA.GOV.UK, <http://data.gov.uk/dataset/uk-civil-service-high-earners>; *Where Does Europe's Money Go? A Guide to EU Budget Data Sources*, OPEN KNOWLEDGE BLOG (July 2, 2015), <http://blog.okfn.org/2015/07/02/where-does-europes-money-go>.

43. G8 OPEN DATA CHARTER, *supra* note 1, ¶¶ 4–5.

44. *Id.*

45. Open Data Toolkit, WORLD BANK, <http://opendatatoolkit.worldbank.org/en/starting.html>.

46. This language is translated from "L'ouverture et le partage des données, c'est la manière, pour un Etat moderne, de s'organiser afin de rendre des comptes, d'ouvrir le dialogue, et de faire confiance à l'intelligence collective des citoyens." *SÉCRETARIAT GÉNÉRAL POUR LA MODERNISATION DE LA FONCTION PUBLIQUE, VADE-MECUM: SUR L'OUVERTURE ET LE PARTAGE DES DONNÉES PUBLIQUES 5* (2013).

accountability and promotes informed participation by the public.<sup>47</sup> A basic consideration of policy for the management of U.S. federal information is that public disclosure of government information is essential to the operation of a democracy.<sup>48</sup> Similarly, the E.U. Public Sector Information Directive says that publishing documents held by the public sector “is a fundamental instrument for extending the right to knowledge, which is a basic principle of democracy.”<sup>49</sup>

The idea of open government is tied to the ideal of transparency of governments’ decisions and activities. Transparency is widely regarded as a precondition for the effective exercise of political rights and freedoms, and for ensuring accountable public authorities.<sup>50</sup> Access to information is a key aspect of democratic institutions that are based on representation, delegation, and accountability. Assessing, debating, and sanctioning public sector behavior requires accurate information.<sup>51</sup> In sum, the proactive disclosure of government data to the public for the purposes of political transparency, accountability and participation is becoming a central tenet in democratic governance.

---

47. OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB MEMORANDUM M-10-06, OPEN GOVERNMENT DIRECTIVE (2009), [https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda\\_2010/m10-06.pdf](https://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf).

48. See OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB CIRCULAR NO. A-130 REVISED, TRANSMITTAL 2, MANAGEMENT OF FEDERAL INFORMATION RESOURCES (1994) (older revision of the Circular); OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, OMB CIRCULAR NO. A-130 REVISED, TRANSMITTAL 4, MANAGEMENT OF FEDERAL INFORMATION RESOURCES (2000) (current version of the Circular). The Circular has a residual role: it does not affect disclosure duties or rights to information under FOIA.

49. Recital 16 of the PSI Directive states:

Making public all generally available documents held by the public sector—concerning not only the political process but also the legal and administrative process—is a fundamental instrument for extending the right to knowledge, which is a basic principle of democracy. This objective is applicable to institutions at every level, be it local, national or international.

Directive 2003/98/EC, *supra* note 35.

50. For an expanded discussion of transparency, see CHRISTOPHER HOOD, & DAVID HEALD, *TRANSPARENCY: THE KEY TO BETTER GOVERNANCE?* (2006); MARK BOVENS ET AL., *THE OXFORD HANDBOOK OF PUBLIC ACCOUNTABILITY* (2014).

51. Like transparency, accountability is a multifaceted concept. For a discussion of dimensions in relation to democracy, see Gijs Jan Brandsma & Thomas Schillemans, *The Accountability Cube: Measuring Accountability*, 23 J. PUB. ADMIN. RES. THEORY 953 (2013).

### 3. *Public Sector Efficiency and Service Delivery*

A third set of pro open data arguments focuses on efficiency: open data should help to save resources and improve public services. For instance, the European Commission says open data will improve health services and traffic management, and help tackle environmental challenges, for instance through monitoring energy consumption.<sup>52</sup>

At the national level, an increasingly popular strategy is to publish performance data of publicly funded organizations.<sup>53</sup> Disclosing inspection and other data is alleged to improve performance of recipients of tax monies, like schools (test scores) and hospitals (deaths, waiting times).<sup>54</sup> Citizens in their capacity as customers are presumed to make better-informed choices when provided with such performance data.<sup>55</sup> Other initiatives serve to improve compliance and to assist in better policymaking or prioritizing enforcement, for instance in the area of food safety standards or building safety.<sup>56</sup> Some open government data initiatives propose a more active role for the public: as an army of armchair auditors who can help identify possible savings.<sup>57</sup>

Furthermore, open data are expected to help public sector bodies carry out their tasks. Many users of open data portals are from the public sector.<sup>58</sup> Efficiency gains made when more transparency about information resources leads to less duplication of information collection, and hence more shared use of resources, are said to improve public sector services.<sup>59</sup> Furthermore, public sector bodies are expected to improve their services when they have more information at their disposal.<sup>60</sup> Efficient use of

52. *Communication from the Commission to the European Parliament et al. on Open Data: An Engine for Innovation, Growth and Transparent Governance*, at 3, COM (2011) 882 final (Dec. 12, 2011), <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0882:FIN:EN:PDF>.

53. Mireille van Eechoud, Inaugural Lecture at the Institute for Information Law at University of Amsterdam: *De Lokroep van Open Data* 9 (May 23, 2014), <http://www.ivir.nl/publicaties/download/1407>.

54. *Id.* at 9.

55. MCKINSEY & CO., *supra* note 32, at 83–85.

56. *See, e.g.*, Michael Flowers, *Beyond Open Data: The Data-Driven City*, in *BEYOND TRANSPARENCY* 185 (Brett Goldstein & Lauren Dyson eds., 2013),

57. *See* BEN WORTHY, *DAVID CAMERON'S TRANSPARENCY REVOLUTION?* 9 (2013), <http://doi.org/10.2139/ssrn.2361428>.

58. *See* WORLD BANK, *supra* note 27.

59. McKinsey, *supra* note 32, at 57–58 (making this case for the energy sector).

60. *See* Alan Feuer, *The Mayor's Geek Squad*, *N.Y. TIMES* (Mar. 23, 2013), <http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html> (discussing the advantages of combining existing data to yield useful information for, e.g., disaster relief efforts or environmental pollution).

information resources is not a new concern of governments. For several decades information management policies have been argued to increase government efficiency.<sup>61</sup>

In an empirical mapping study, we found that different arguments for open data obtain varying levels of attention amongst different actors in different forms of digital media.<sup>62</sup> For example, in English language mainstream media outlets arguments and examples about the economic growth and technological innovation potential of open data received more attention than those related to public participation or democratic accountability. On social media platforms such as Twitter, distinct groups of actors were interested in different sets of topics around open data such that, for example, some were interested in startups and smart cities, and others were interested in transparency and open government.<sup>63</sup>

In sum, open data policies serve diverse interests. For the purposes of this Article, these can be clustered into: (1) innovation and economic growth, (2) political accountability and democratic participation, and (3) public sector efficiency.

#### B. PRIVACY INTERESTS

At the global level, the right to privacy is protected under, for instance, the United Nations Declaration of Human Rights<sup>64</sup> and the International Covenant on Civil and Political Rights.<sup>65</sup> In the United States, the Fourth Amendment and other laws protect privacy.<sup>66</sup> In Europe, the European Convention on Human Rights,<sup>67</sup> the European Union Charter of

---

61. FEDERAL INFORMATION DISSEMINATION IN AN ELECTRONIC AGE, *supra* note 37.

62. Jonathan Gray et al., Mapping the Politics of Open Data on Digital Media (in preparation) (on file with authors).

63. *Id.*

64. Universal Declaration of Human Rights, art. 12, G.A. Res. 217A (III), U.N. Doc. A/810, at 71 (1948).

65. International Covenant on Civil and Political Rights, art. 17, Dec. 16, 1966, S. Treaty Doc. No. 95-20, 6 I.L.M. 368 (1967), 999 U.N.T.S. 171.

66. *See generally* WILLIAM CUDDIHY, THE FOURTH AMENDMENT: ORIGINS AND ORIGINAL MEANING 602–1791 (2009); DANIEL SOLOVE & PAUL SCHWARTZ, INFORMATION PRIVACY LAW 260–335 (5th ed., 2014).

67. Convention for the Protection of Human Rights and Fundamental Freedoms, art. 8, Nov. 4, 1950, 213 U.N.T.S. 222 [hereinafter ECHR] (also referred to as the European Convention on Human Rights).

Fundamental Rights,<sup>68</sup> national constitutions, and other laws protect privacy.<sup>69</sup>

Public sector bodies hold an enormous amount of personal information, and this amount will likely grow. For instance, so-called “smart cities” may provide the public sector information about people such as up-to-date location data of cars, and detailed electricity metering data.<sup>70</sup> And, as public sector bodies offer more services online, they will obtain even more information about people.<sup>71</sup> Sometimes citizens volunteer personal information, for example when they use public services. But public authorities can also collect information through third parties, like educational and health care institutions.<sup>72</sup> And authorities can compel citizens to provide personal information. This element of force heightens privacy concerns.

We distinguish three broad categories of privacy concerns regarding open data: (1) the chilling effects on people in their interaction with the public sector, (2) a lack of individual control over personal information, and (3) the use of open data as input for social sorting and discriminatory practices.<sup>73</sup>

---

68. Charter of Fundamental Rights of the European Union of the European Parliament, arts. 7–8, 2010 O.J. (C 83) 2, 1 [hereinafter E.U. Charter of Fundamental Rights].

69. See, e.g., Grondwet voor het Koninkrijk der Nederlanden [Constitution of the Kingdom of the Netherlands], art. 10. Furthermore, each E.U. member state has a national data protection act implementing the European Parliament’s directive “on the protection of individuals with regard to the processing of personal data and on the free movement of such data.” Council Directive 95/46/EC, art. 28, 1995 O.J. (L 281).

70. A smart city has been defined “as one that has digital technology embedded across all city functions.” *Definitions and Overviews*, SMART CITIES COUNCIL, <http://smarcitiescouncil.com/smart-cities-information-center/definitions-and-overviews> (last visited May 1, 2015); see also Robert G. Hollands, *Will the Real Smart City Please Stand Up? Intelligent, Progressive or Entrepreneurial?*, 12 CITY 303 (2008).

71. Teresa Scassa, *Privacy and Open Government*, 6 FUTURE INTERNET 397, 397–98 (2014).

72. See Solove, *Access and Aggregation*, *supra* note 7, at 1142–50 (contains an overview of federal, state, and local record collection in the United States).

73. FREDERIK ZUIDERVEEN BORGESIU, *IMPROVING PRIVACY PROTECTION IN THE AREA OF BEHAVIOURAL TARGETING* 53–63 (2015). The three categories are based on that study, which does not concern open data. In this Article, we adapt the categories to the open data context.

### 1. *Chilling Effects*

First, a chilling effect can occur if people interacting with public bodies fear that their information will be stored, or will be made public.<sup>74</sup> For example, people might be less inclined to contact public sector agencies if they doubt that their personal data will remain confidential.<sup>75</sup>

People often provide personal information when engaging with public sector bodies. Public sector bodies often require information, for example, when people apply for a planning permission or business license, attempt to comply with health and safety standards, or submit tax claims or grant applications. The collection, use and exchange of personal information are part of the normal fabric of public sector activity. Many public services cannot be delivered without these activities.

People might refrain from contacting the public sector if they fear their personal information will not be kept confidential. Especially people with questions about diseases, pregnancies, drugs, financial troubles, or suicidal thoughts might refrain from asking for help. Jeff Jonas and Jim Harper illustrate the importance of communicating with the public sector without disclosing too much personal information with an example regarding a migrant.<sup>76</sup> Say Alice is a migrant who thinks her residence permit contains errors. If she thinks that visiting the immigration website will bring her to the attention of immigration law enforcement, she might forego looking for information. “If she cannot communicate this information anonymously, she almost certainly will not ask questions or volunteer information, denying herself help she might deserve while denying policymakers relevant information.”<sup>77</sup> If Alice thought her data would be disclosed to others in and outside government, such a chilling effect might be greater.

By itself the chilling effect already harms the individual who refrains from an activity she might otherwise engage in. But if somebody does not seek help because of a chilling effect, for instance if someone does not seek information regarding a disease, he or she may also experience more

---

74. See KIERON O'HARA, *TRANSPARENT GOVERNMENT, NOT TRANSPARENT CITIZENS: A REPORT ON PRIVACY AND TRANSPARENCY FOR THE CABINET OFFICE* 24 (2011), [http://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/61279/transparency-and-privacy-review-annex-a.pdf](http://www.gov.uk/government/uploads/system/uploads/attachment_data/file/61279/transparency-and-privacy-review-annex-a.pdf).

75. Jeff Jonas & Jim Harper, *Open Government: The Privacy Imperative*, in *OPEN GOVERNMENT: TRANSPARENCY, COLLABORATION, AND PARTICIPATION IN PRACTICE* 315, 322–23 (Daniel Lathrop & Laurel R. T. Ruma eds., 2010).

76. *Id.*

77. *Id.* at 317.

tangible harms. People foregoing treatment of infectious diseases could harm society as a whole.

Uncertainty about what happens with one's personal information can ultimately adversely impact the quality of public services. As Teresa Scassa notes, with open data "there is a risk not only to individual privacy, but also to the relationship of trust that is meant to exist between citizens and their government."<sup>78</sup> Government statistics offices have realized for a long time that confidentiality of census answers is important—otherwise people might not give honest answers anymore. Trust in public authorities could diminish if people do not believe that their personal data will remain confidential.<sup>79</sup> In sum, open data policy could lead to a chilling effect on people communicating with the public sector, which is a privacy problem.

## 2. *Lack of Control over Personal Information*

A second privacy concern is that people lack control over their personal information if that information is released as open data. Publicly releasing personal information as open data can be especially troublesome because open data policy in its most liberal form implies that unlimited numbers of re-users can use the data for any purpose.

Many privacy definitions focus on individual control over personal information. For instance, Alan Westin defined privacy in 1967 as "the claim of individuals, groups, or institutions to determine when, how, and to what extent information about them is communicated to others."<sup>80</sup> Many scholars use similar privacy definitions.<sup>81</sup> The privacy as control perspective is apparent in legal practice. For instance, the U.S. Supreme Court has described privacy as "the individual's control of information concerning his or her person."<sup>82</sup> The German Supreme Court says a person has, in principle, the right "to determine for himself whether his

78. Scassa, *supra* note 71, at 408.

79. *See, e.g.*, U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-01-126SP, RECORD LINKAGE AND PRIVACY: ISSUES IN CREATING NEW FEDERAL RESEARCH AND STATISTICAL INFORMATION 18 (2001).

80. ALAN F. WESTIN, *PRIVACY AND FREEDOM* 7 (reprint 1970) (1967).

81. *See* Charles Fried, *Privacy*, 77 *YALE L.J.* 475, 482 (1968) (discussing that privacy "is not simply an absence of information about us in the minds of others; rather it is the *control* we have over information about ourselves."). *See also* A.R. MILLER, *THE ASSAULT ON PRIVACY: COMPUTERS, DATA BANKS, AND DOSSIERS* 25 (1971) (describing privacy as "the individual's ability to control the circulation of information relating to him").

82. *U.S. Dep't of Justice v. Reporters Comm. for Freedom of the Press*, 489 U.S. 749, 763 (1988).

personal data should be divulged or utilized.”<sup>83</sup> Privacy as control has deeply influenced the Fair Information Principles.<sup>84</sup> The privacy as control perspective does not capture all the subtleties of privacy. Nevertheless, a loss of individual control over personal information is widely seen as a privacy problem.<sup>85</sup>

A lack of individual control over personal information can lead to subjective and objective privacy harm. Objective harm is, in Ryan Calo’s words, “the unanticipated or coerced use of information concerning a person against that person.”<sup>86</sup> The Eightmaps website provides an example of objective harm resulting from data released by the public sector.<sup>87</sup> Proposition 8 was a 2008 proposal to amend the California constitution with a referendum to ban gay marriage.<sup>88</sup> California law requires that campaign donations be published.<sup>89</sup> An anonymous website publisher took information regarding donors who supported Proposition 8, and overlaid that information on Google Maps.<sup>90</sup> The map showed information such as the donor’s name, approximate location, and the amount donated. Some of the donors received death threats, or were the victim of boycotts.<sup>91</sup> The dissemination of correct information can already produce objective harms, but the potential of harm arising from the public release of inaccurate or false data is at least as large.

---

83. Bundesverfassungsgericht [BVerfG] [Federal Constitutional Court] Mar. 25, 1982, BGBl. I 369, 1982 (Ger.), *translated in* E.H. Riedel, *New Bearings in German Data Protection*, 5 HUM. RTS. L.J. 94, 101 (1984).

84. *See, e.g.*, COLIN J. BENNETT, *REGULATING PRIVACY: DATA PROTECTION AND PUBLIC POLICY IN EUROPE AND THE UNITED STATES* 14 (1992). *See infra* Part IV.

85. *See, e.g.*, Fahriye Seda Gürses, *Multilateral Privacy Requirements Analysis in Online Social Networks* (May 2010) (Ph.D. thesis, University of Leuven); HELEN NISSENBAUM, *PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE* (2010); Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477 (2006).

86. Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L.J. 1131, 1133 (2011).

87. *Eightmaps.com and Too Much Information*, DALLAS MORNING NEWS (Jan. 14, 2009); *see also* Michael Shin, *Show Me the Money! The Geography of Contributions to California’s Proposition 8*, 1 CAL. J. POL. & POL’Y 10 (2009). *See generally* on privacy-invasive online map services: Mark Burdon, *Privacy Invasive Geo-Mashups: Privacy 2.0 and the Limits of First Generation Information Privacy Laws*, 2010 U. ILL. J.L. TECH. & POL’Y 1 (2010).

88. CAL. CONST. art. I, § 7.5 (enacted as California 2008 Ballot Proposition 8), *ruled unconstitutional in* *Perry v. Schwarzenegger*, 704 F.Supp.2d 921 (N.D. Cal. 2010).

89. Deborah G. Johnson, Priscilla M. Regan & Kent Wayland, *Campaign Disclosure, Privacy and Transparency*, 19 WM. & MARY BILL RTS. J. 959, 972 (2011).

90. *Id.*

91. *Id.* *See also* Brad Stone, *Disclosure, Magnified on the Web*, N.Y. TIMES, Feb. 7, 2009, at BU3, <http://www.nytimes.com/2009/02/08/business/08stream.html>.

The feeling of having no control over one's personal information is a "subjective harm," described by Calo as "the perception of loss of control that results in fear or discomfort."<sup>92</sup> Many people are uncomfortable with organizations processing large amounts of information about them. Furthermore, there is often information asymmetry between the individual and the organization that uses personal information. People may know that information about them is collected and stored, but may not know how this will be used. If people do not know who holds data about them, they cannot exercise control over those data.<sup>93</sup> Releasing data to an undetermined number of re-users aggravates the lack of control.

Furthermore, data privacy rules that apply to the public sector are often stricter than those that apply to the private sector.<sup>94</sup> However, if the public sector releases personal data as open data, that is, with no restrictions, the private sector can subsequently use those data, subject to more lenient (statutory) rules.<sup>95</sup> Hence, releasing personal data as open data reduces privacy protection. Furthermore, the more datasets governments disclose, the richer the possibilities for re-identification. In sum, releasing personal information as open data causes a lack of individual control over personal information.

### 3. *Social Sorting and Discrimination*

A third privacy-related concern is that open data could be used as input for social sorting and discriminatory practices.<sup>96</sup> For instance, if the public sector released personal data, data brokers would likely be among

92. Calo, *supra* note 86, at 1143.

93. See generally Alessandro Acquisti & Jens Grossklags, *What Can Behavioral Economics Teach Us About Privacy?*, in DIGITAL PRIVACY: THEORY, TECHNOLOGIES AND PRACTICES 363 (Sabrina De Capitani di Vimercati et al. eds., 2007); ZUIDERVEEN BORGESIU, IMPROVING PRIVACY PROTECTION, *supra* note 73, at 201–05. According to Solove, the feeling of lost control resembles Franz Kafka's THE TRIAL. DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 38 (2004). He suggests the main problem is "not knowing what is happening, having no say or ability to exercise meaningful control over the process." *Id.* at 38.

94. For instance, in the United States the Privacy Act of 1974 does not apply to the private sector. Pub. L. No. 93-579, 88 Stat. 1896 (codified at 5 U.S.C. § 552a (2012)). In the European Union, firms more easily meet the required legal basis test for personal data processing than public sector bodies do. See Directive 95/46/EC, *supra* note 69. Article 7(f) applies to firms; Article 7(e) applies to the public sector.

95. Scassa, *supra* note 71, at 405, 402.

96. See, e.g., Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. (forthcoming 2016), <http://ssrn.com/abstract=2477899>; EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 51 (2014), [http://www.whitehouse.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf); Scassa, *supra* note 71, at 407.

the main re-users.<sup>97</sup> Data brokers are “companies that collect consumers’ personal information and resell or share that information with others.”<sup>98</sup> The information can be used, for instance, for direct marketing, credit scoring, or screening job applicants.<sup>99</sup>

Many find data brokers’ activities unfair and privacy-invasive.<sup>100</sup> As the Federal Trade Commission notes, personal information could be used for unfair discrimination. For instance, a company might use the information that there is a “Smoker in Household” to conclude that people in that household should not be offered insurance.<sup>101</sup> In surveillance studies, such practices are called “social sorting.” As David Lyon explains, social sorting involves “obtain[ing] personal and group data in order to classify people and populations according to varying criteria, to determine who should be targeted for special treatment, suspicion, eligibility, inclusion, access, and so on.”<sup>102</sup> Social sorting is not inherently bad or good.<sup>103</sup>

For social sorting, data brokers can also use open data that do not include personal information. For instance, the average housing price in a certain zip code is not personal information. But that average price could be matched with somebody’s address to estimate the value of his or her house. Hence, non-personal information can be used to enrich digital dossiers about people.

The following is another example of a social sorting effect resulting from open data. Suppose a city council releases crime statistics. A vendor

---

97. Thomas P. Keenan, *Are They Making Our Privates Public? Emerging Risks of Governmental Open Data Initiatives*, in *PRIVACY AND IDENTITY MANAGEMENT FOR LIFE 1*, 11 (Jan Camenisch et al. eds., 2012). *See also* Solove, *Access and Aggregation*, *supra* note 7, at 1148–50.

98. *See also* U.S. FEDERAL TRADE COMMISSION, *DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY 1* (2014), <http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf>.

99. *See* Scassa, *supra* note 71, at 407.

100. *See generally* Chris Jay Hoofnagle, *Big Brother’s Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement*, 29 N.C. J. INT’L L. & COM. REG. 595 (2003); JOSEPH TUROW, *NICHE ENVY: MARKETING DISCRIMINATION IN THE DIGITAL AGE* (2006); JOSEPH TUROW, *THE DAILY YOU: HOW THE NEW ADVERTISING INDUSTRY IS DEFINING YOUR IDENTITY AND YOUR WORTH* (2011).

101. FEDERAL TRADE COMMISSION, *DATA BROKERS*, *supra* note 98, at 55–56.

102. David Lyon, *Surveillance as Social Sorting: Computer Codes and Mobile Bodies*, in *SURVEILLANCE AS SOCIAL SORTING: PRIVACY, RISK AND DIGITAL DISCRIMINATION 13* (David Lyon ed., 2002) [hereinafter Lyon, *Surveillance as Social Sorting*].

103. David Lyon, Kevin Haggerty & Kirstie Ball, *Introducing Surveillance Studies*, in *ROUTLEDGE HANDBOOK OF SURVEILLANCE STUDIES 3* (Kirstie Ball, Kevin Haggerty & David Lyon eds., 2012).

of GPS car systems can overlay its own maps with the crime data in order to designate high-crime areas. The car GPS system can then route the driver around those areas. The practice could be seen as unfair for the people and businesses in that newly invented no-go area. In the no-go areas, insurance premiums might rise, and real estate prices and shop profits might drop.

In sum, potential privacy problems regarding open data include chilling effects on people communicating with the public sector, a lack of individual control over personal information, and discriminatory practices enabled by the released data. Hence, especially when datasets contain personal data, public sector bodies should give due consideration to the risks of disclosing data. We discuss below how to strike a balance between open data policy and privacy. But first we turn to the rules and guidelines that govern the disclosure of public sector information.

### III. GOVERNANCE OF PUBLIC SECTOR INFORMATION

In this section we discuss governance frameworks regarding access to public sector information. We discuss norms that govern open data, and norms that govern access to public sector information more generally. Freedom of information laws provide inspiration on how to strike a balance between privacy and transparency in the open data context.

#### A. OPEN DATA NORMS

Obligations for public authorities to release information as open data tend not to be encoded in hard law. Rather, open data policy is often promoted through administrative hierarchies, whereby the policy objectives, targets, and instructions range from superficial and permissive to detailed and strict.<sup>104</sup> Open data policymaking is partly shaped through

---

104. For example, the 2013 order by President Obama breathes ambition and decisiveness, and the elaboration by the Office for Management & Budget of its Open Data Policy Memorandum contains specific duties for departments to create lists of available data sets (“Public Data Listing”) and engage with user groups to prioritize release. OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, SUPPLEMENTAL GUIDANCE ON THE IMPLEMENTATION OF MEMORANDUM M-13-13 (2013), <https://project-open-data.cio.gov/implementation-guide/>; see also OMB MEMORANDUM M-13-13, OPEN DATA POLICY, *supra* note 25. The E.U.’s Public Sector Information Directive shows a preference for the release of data in open formats, and also demands that member states make practical arrangements “that help re-users in their search for documents available for re-use,” e.g., in the form of asset registers. Directive 2003/98/EC, *supra* note 35. The E.C. Guidelines clearly favor pro-active release of data as open.

political commitment in international forums such as the G8 and the Open Government Partnership.<sup>105</sup>

Open data initiatives rely on norms that regulate access to information. After all, open government data are, by definition, publicly available data. A myriad of such norms exists at the national level. The most generic disclosure duties arise under freedom of information acts, which typically cover the executive branch. Constitutional and administrative norms that help cement basic checks and balances also have implications for access to information, mandating for example that legislative texts are published,<sup>106</sup> and that the public has access to court decisions.<sup>107</sup>

Additionally, many countries have dedicated laws that govern information production for specific purposes, such as (national) statistics to aid policy development and monitoring,<sup>108</sup> land registries to facilitate secure property transactions, business registers,<sup>109</sup> or earth observation data produced for environmental and agricultural management.<sup>110</sup> Such specific laws will often lay down modalities for access. For example, confidentiality of identifiable information is of fundamental interest for the production of reliable and useful statistics. Hence, a basic principle in instruments that govern the production and dissemination of statistics is that personal information supplied for statistical purposes will not be disclosed or used

---

105. Through the *Open Data Charter*, the members of G8 have committed to drafting national open data action plans. *Supra* note 1. The same mechanism is used by the *Open Government Partnership*. *Supra* note 6.

106. *See, e.g.*, 1958 CONST. arts. 10–11 (Fr.); GRUNDGESETZ FÜR DIE BUNDESREPUBLIK DEUTSCHLAND [GRUNDGESETZ] [GG] [BASIC LAW], May 23, 1949, BGBl. I, Art. 82 (Ger.).

107. For example, Article 6 of the European Convention on Human Rights (on the right to a fair trial) prescribes that court decisions are to be made public. ECHR, *supra* note 67, art. 6. This will usually be through delivery in court but may be achieved by other means as well. *See* COUNCIL OF EUROPE, GUIDE ON ARTICLE 6, at 49–50 (2013). That a right to information is no guarantee for easy and affordable access is witnessed by the electronic access system for federal courts. *See* Vera Eidelman & Amul Kalia, *Right to Know: The PACER Mess and How to Clean It*, ELEC. FRONTIER FOUND. (Sept. 2, 2014), <https://www.eff.org/deeplinks/2014/09/right-know-pacer-mess-and-how-clean-it>.

108. *See, e.g.*, Stb. 2003, p. 551 [Act on the Central Bureau of Statistics] (Neth.); Statistics Act, R.S.C. 1985, c. S-19 (Can.); Statistics and Registration Service Act, 2007, c. 18 (U.K.).

109. *See, e.g.*, Stb. 2007, p. 153 [Act on Trade Register] (Neth.); Companies Act, 2006, c. 46 (U.K.); Handelsgesetzbuch [Act on Trade Register], Oct. 23, 2008, BGBl. III at 4100-1, § 8 (Ger.).

110. Mireille van Eechoud, *Commercialization of Public Sector Information: Delineating the Issues*, in THE FUTURE OF THE PUBLIC DOMAIN: IDENTIFYING THE COMMONS IN INFORMATION LAW 279, 281–83 (Lucie Guibault & P. Bernt Hugenholtz eds., 2006).

for other (administrative) purposes.<sup>111</sup> In the interest of research, some statistics offices organize secure environments, where researchers can access micro-data under strict conditions. While no international legal right to (re)use public sector information exists, access to government information is increasingly recognized as a human right.<sup>112</sup>

#### B. ACCESS TO INFORMATION NORMS

Several international courts see access rights as part of, or closely connected to, the right to freedom of expression.<sup>113</sup> However, access rights are also recognized in case law of the European Court of Human Rights in

---

111. *See* CONFERENCE OF EUROPEAN STATISTICIANS, FUNDAMENTAL PRINCIPLES OF OFFICIAL STATISTICS (1991). They were since updated and endorsed by the U.N. General Assembly. G.A. Res. 68/261, U.N. Doc. A/68/261 (Jan. 29, 2014). Principle 6 reads: “Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.” *Id.* There are similar examples at the national level. *See, e.g.*, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), Pub. L. 107-347, 116 Stat. 2962 (2002); Stb. 2003, p. 516 [Dutch Statistics Act] (Neth.).

112. For an extensive analysis of different human rights based conceptualizations of access to government information, see CHERYL A. BISHOP, ACCESS TO INFORMATION AS A HUMAN RIGHT (2011).

113. A right to access of government information is guaranteed under Article 13 (Freedom of Thought and Expression) of the American Convention on Human Rights, and states have a positive obligation to provide access, subject only to access restrictions that are proportionate and for reasons permitted by the Convention. *Claude-Reyes et al. v. Chile*, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 151, ¶ 77 (Sept. 9, 2006). Refusal to grant access to government information to a public watchdog violates the right to freedom of expression under Article 10 of the ECHR. *Youth Initiative for Human Rights v. Serbia*, App. No. 48135/06, Judgment, 2013 Eur. Ct. H.R. 584 (2013); ECHR, *supra* note 67, art. 10. In another case, the European Court of Human Rights (ECtHR) conceded it “has recently advanced towards a broader interpretation of the notion of ‘freedom to receive information’ and thereby towards the recognition of a right of access to information.” *TASZ v. Hungary*, App. No. 37374/05, 2009 Eur. Ct. H.R., ¶ 35 (2009). Previously it had rejected the claim that ECHR Article 10 includes a right to access government information, or a positive obligation for states to collect and disseminate information. *See, e.g.*, *Guerra v. Italy*, App. No. 14967/89, 1998 Eur. Ct. H.R. 7 (1998).

the context of the right to private life.<sup>114</sup> By contrast, access rights may be conceived of as stand-alone constitutional rights.<sup>115</sup>

The Tromsø Convention of the Council of Europe concerns access to government information,<sup>116</sup> but it is unlikely that enough member states will ratify this convention for it to enter into force any time soon.<sup>117</sup> Much more successful is the U.N. Aarhus Convention of 1998, with nearly fifty contracting states.<sup>118</sup> The Aarhus Convention provides for a right of access to environmental information as part of every citizen's right to an adequate environment and duty to safeguard the environment for future generations.<sup>119</sup>

A fundamental right of access does not necessarily imply that authorities must actively disclose information to the general public in electronic form without use-restrictions. But open government agendas do steer policy in that direction. At the global level, the Open Government Partnership promotes proactive disclosure in reusable formats.<sup>120</sup>

In various human rights domains, proactive disclosure is also advocated. The U.N. rapporteur on Human Rights typifies the right to access government information as "one of the central components of the right to freedom of opinion and expression."<sup>121</sup> To give effect to the right

---

114. The ECtHR recognized a duty to impart information for the government as part of the right to respect for private life (under Article 8 of the ECHR) on various occasions: where it concerned access to foster care records, *Gaskin v. UK*, App. No. 10454/83, 12 Eur. H.R. Rep. 36 (1989), and with respect to information about environmental pollution (threatening citizens' health), *Guerra v. Italy*, *supra* note 113; *Onderyildiz v. Turkey*, 2004-XII Eur. Ct. H.R. 81. In these cases, applicants had a special interest.

115. For example, Article 42 of the Charter of Fundamental Rights of the European Union provides that any citizen of the Union has a right of access to documents held by E.U. institutions. E.U. Charter of Fundamental Rights, *supra* note 68, art. 42. For an in depth analysis of access rights of a wider openness agenda, see Alberto Alemanno, *Unpacking the Principle of Openness in EU Law: Transparency, Participation and Democracy*, 39 EUR. L. REV. 72 (2014).

116. Tromsø Convention, Council of Europe Convention on Access to Official Documents, *opened for signature* June 18, 2009, C.E.T.S. No. 205 (not yet ratified).

117. Mireille van Eechoud & Katleen Janssen, *Rights of Access to Public Sector Information*, 6 MASARYK U. J.L. & TECH. 471, 486 (2012).

118. Aarhus Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters, *opened for signature* June 25, 1998, 2161 U.N.T.S. 447 (entered into force Oct. 30, 2001).

119. *Id.*

120. See *Open Government Declaration*, OPEN GOVERNMENT PARTNERSHIP, <http://www.opengovpartnership.org/about/open-government-declaration> (last visited May 1, 2015).

121. Special Rapporteur, *Report on the Promotion and Protection of the Right to Freedom of Opinion and Expression in Accordance with Human Rights Council Resolution 16/4*,

of access to information under Article Nineteen of the United Nations International Covenant on Civil and Political Rights and the Universal Declaration of Human Rights, “parties should proactively put in the public domain Government information of public interest” and “make every effort to ensure easy, prompt, effective and practical access to such information.”<sup>122</sup> In 2006, the Inter-American Court of Human Rights held that States have a positive obligation to legislate freedom of information laws or take other measures that ensure access to government information.<sup>123</sup>

The adoption rate of freedom of information laws has accelerated on all continents over the past decade. Today nearly a hundred countries have enacted freedom of information laws.<sup>124</sup> Some freedom of information laws contain provisions on proactive disclosure of information.<sup>125</sup> These tend to be vague and rather limited in scope. Traditionally access laws focus on disclosure of information on request by a member of the public. Access laws detail how requests can be made and how decisions must be reached.<sup>126</sup> A basic principle in freedom of information acts is that citizens do not have to motivate why they want access; the public interest in

---

*transmitted by Note of the Secretary-General*, U.N. Doc. A/68/362 (Sept. 4, 2013). *See also* Human Rights Council Res. 12/12, Right to the Truth, 12th Sess., Oct. 1, 2009, U.N. GAOR, 68th Sess., A/HRC/RES/12/12, at 3 (Oct. 12, 2009) (“the public and individuals are entitled to have access, to the fullest extent practicable, to information regarding the actions and decision-making processes of their Government, within the framework of each State’s domestic legal system”); Inter-American Commission on Human Rights [IACHR], *The Right to Truth in the Americas*, IACHR Doc. OEA/Ser.L/V/II.152 (Aug. 13, 2014).

122. Human Rights Comm. on Article 19: Freedoms of Opinion and Expression, General Comment No. 34, Rep. on its 102d Sess., July 11–29, 2011, U.N. Doc. CCPR/C/GC/34, ¶ 19 (Sept. 12, 2011).

123. *Claude-Reyes et al. v. Chile, Merits, Reparations, and Costs, Judgment*, Inter-Am. Ct. H.R. (ser. C) No. 151, ¶¶ 77, 102 (Sept. 9, 2006).

124. *See* Map, GLOBAL RIGHT TO INFORMATION RATING, <http://www.rti-rating.org>.

125. For an analysis of the drivers of pro-active disclosure of government information and its growing enactment in binding norms, see Helen Darbshire, *Proactive Transparency: The Future of the Right to Information?* (World Bank Inst. Governance Working Paper Series, No. 56,598, 2010).

126. *See* JONATHAN GRAY & HELEN DARBISHIRE, BEYOND ACCESS: OPEN GOVERNMENT DATA & THE RIGHT TO (RE)USE PUBLIC INFORMATION (Creative Commons, 2011); Mireille van Eechoud et al., *Good Practices Collection on Access to Data*, LAPSI (July 11, 2014) [hereinafter *Good Practices Collection*].

disclosure is considered a given.<sup>127</sup> A right to access information does not necessarily imply that the information can subsequently be used freely.<sup>128</sup>

Generally, freedom of information laws do not prescribe how data must be made available (for example in an open format, machine readable, with a certain frequency).<sup>129</sup> Usually, information disclosed under freedom of information laws is not required to be legally or technically open.<sup>130</sup> It is, however, a common feature that public bodies must, wherever possible, respect the mode of supply preferred by the requesting party, if the documents are available in such form or easily so produced.<sup>131</sup> Freedom of information laws usually contain privacy provisions, as discussed next.

### C. ACCESS TO INFORMATION NORMS AND PRIVACY

Machine readable, bulk-downloadable open data complicate a problem that was already a difficult one in the pre-digital era. Since at least the 1970s, countries have grappled with the problem of balancing privacy protection and public sector transparency.<sup>132</sup> Generic freedom of information laws typically aim to accommodate privacy interests, for example by reserving access to personal information to parties with particular interests, or by only making records available in secure reading rooms.

Two balancing models regarding privacy and transparency can be distinguished in freedom of information laws. First, sometimes privacy is an absolute limitation to disclosure. That is, the legislator has done the balancing ex-ante. For example, the Dutch Freedom of Information Act provides that certain types of sensitive personal data (for example data concerning medical matters or religion) may never be disclosed.<sup>133</sup>

---

127. GRAY & DARBISHIRE, *supra* note 126; *Good Practices Collection*, *supra* note 126.

128. For instance, before implementation of the E.U. Public Sector Information Directive, the Belgian federal freedom of information act stipulated that no commercial use was allowed of information obtained under the act. *See* Wet betreffende de openbaarheid van bestuur of Apr. 11, 1994, BELGISCH STAATSBLAD [B.S.] [Official Gazette of Belgium], June 30, 1994 (deleted by Act N. 2007-1600, Mar. 7, 2007).

129. *See* the analysis of over forty freedom of information acts, GRAY & DARBISHIRE, *supra* note 126; *Good Practices Collection*, *supra* note 126.

130. GRAY & DARBISHIRE, *supra* note 126.

131. *See, e.g.*, Aarhus Convention of 1998, *supra* note 118, art. 4.

132. For instance, in 1973 Sweden adopted its data privacy law partly to ensure that the generous Swedish regime for access to official documents, which dates back to 1776, would not unduly interfere with privacy. *See* GLORIA GONZÁLEZ FUSTER, THE EMERGENCE OF PERSONAL DATA PROTECTION AS A FUNDAMENTAL RIGHT OF THE EU 59 (2014).

133. Wet openbaarheid van bestuur [Dutch Freedom of Information Act], Stb. 1991, art. 10(1)d (Neth.).

Second, sometimes freedom of information laws include a relative privacy exemption, to be weighed against the public interest in disclosure on a case-by-case basis.<sup>134</sup> U.S. freedom of information law exempts disclosure of personal, medical and similar files.<sup>135</sup> The test is whether disclosure “would constitute a clearly unwarranted invasion of personal privacy.”<sup>136</sup> Personal information gathered as part of law enforcement is also exempt, if disclosure “could reasonably be expected to constitute an unwarranted invasion of personal privacy.”<sup>137</sup> If privacy interests prevent disclosure, it is common for freedom of information laws to demand that exempted information is redacted so that the remainder can be released, even if cleaning documents is labor intensive.<sup>138</sup>

The regulation that governs access to documents from E.U. institutions (such as the Council of Ministers, Parliament, and Commission) stipulates that access to a document shall be refused if “disclosure would undermine the protection of privacy . . . in particular in accordance with Community legislation regarding the protection of personal data.”<sup>139</sup> The Obama Freedom of Information Memorandum states: “In the face of doubt, openness prevails.”<sup>140</sup>

At global human rights forums, the presumption is that the public interest in access to public sector information (as part of the freedom of expression) trumps privacy and other interests. Human rights rapporteurs for the United Nations argue that access to information should be granted unless disclosure would cause serious harm to a protected interest such as

134. The Dutch Freedom of Information Act provides such a relative ground for non-disclosure, where the public’s right to know does not outweigh a person’s interest to have his or her private sphere protected. *Id.*, art. 10(2)e.

135. 5 U.S.C. § 552(b)(6) (2012) (known as Exemption 6 of Electronic Freedom of Information Act of 1966, Pub. L. No. 104-231, 110 Stat. 3048).

136. *Id.*

137. 5 U.S.C. § 552(b)(7).

138. For example, Article 4(4) of the Aarhus Convention exempts the release of personal data (if confidential under domestic law); Article 4(6) obliges states to redact the documents. Aarhus Convention of 1998, *supra* note 118, art. 4.

139. Council Regulation 1049/2001, art. 4, 2001 O.J. (L 145) 43. The way the institutions have interpreted this limitation is controversial; the European Ombudsman and the European Data Protection Supervisor signal overzealous interpretation of the rules on data protection as a threat to transparency. How the scales tip thus depends as much on the prevailing culture of transparency (or secrecy) as on the black letter. *See* H. R. KRANENBORG, TOEGANG TOT DOCUMENTEN EN BESCHERMING VAN PERSOONSGEGEVENS IN DE EUROPESE UNIE [ACCESS TO DOCUMENTS AND DATA PROTECTION IN THE EUROPEAN UNION] 188–94 (2007).

140. *See* President Barack Obama, Memorandum, Freedom of Information Act, 74 Fed. Reg. 15 (Jan. 21, 2009).

privacy that outweighs the interest in disclosure.<sup>141</sup> The rapporteurs also stress the importance of proactive disclosure obligations, and posit that “access to information law should, to the extent of any inconsistency, prevail over other legislation.”<sup>142</sup>

Particularly for the disclosure in the interest of political accountability and public debate, judgments in which the right to freedom of expression and the right to privacy are balanced can give guidance. The European Court of Human Rights recognizes the importance of proactive release of data on the Internet as a means to ensure effective transparency and accountability. In the *Wytych* case, the Court rejected the claim by an elected local councilor who argued that by requiring him to disclose information on his financial interests online, the Polish legislature infringed his right to privacy under Article 8 of the European Convention on Human Rights. The Court noted “[t]he general public has a legitimate interest in ascertaining that local politics are transparent and Internet access to the declarations makes access to such information effective and easy. Without such access, the obligation would have no practical importance or genuine incidence on the degree to which the public is informed about the political process.”<sup>143</sup>

Earlier, the European Court of Human Rights held that the privacy interests of politicians and higher public officials must yield to access rights.<sup>144</sup> The Court considered “that it would be fatal for freedom of expression in the sphere of politics if public figures could censor the press and public debate in the name of their personality rights, alleging that their opinions on public matters are related to their person and therefore constitute private data which cannot be disclosed without consent.”<sup>145</sup>

---

141. *See, e.g.*, AMBEYI LIGABO, U.N. SPECIAL RAPPORTEUR ON FREEDOM OF OPINION AND EXPRESSION ET AL., JOINT DECLARATION, INTERNATIONAL MECHANISMS FOR PROMOTING FREEDOM OF EXPRESSION (2006) [hereinafter JOINT DECLARATION, PROMOTING FREEDOM OF EXPRESSION (2006)] (signed by the U.N. Special Rapporteur on Freedom of Opinion and Expression, the OSCE Representative on Freedom of the Media, the OAS Special Rapporteur on Freedom of Expression and the African Commission on Human and Peoples’ Rights). *See also* JOINT DECLARATION, INTERNATIONAL MECHANISMS FOR PROMOTING FREEDOM OF EXPRESSION (2004) [hereinafter JOINT DECLARATION, PROMOTING FREEDOM OF EXPRESSION (2004)].

142. *See, e.g.*, JOINT DECLARATION, PROMOTING FREEDOM OF EXPRESSION (2004), *supra* note 141.

143. *Wytych v. Poland*, App. No. 2428/05, 2005 Eur. Ct. H.R. (admissibility decision).

144. *TASZ v. Hungary*, *supra* note 113, ¶ 37.

145. *TASZ v. Hungary*, *supra* note 113, ¶ 37.

In conclusion, regulation and case law regarding freedom of public sector information can provide inspiration on how to strike the balance between privacy and transparency in the open data context. Apart from that, there are more general principles to balance privacy-related interests and other interests. We turn to those Fair Information Principles now.

#### IV. GOVERNANCE OF PERSONAL INFORMATION

The Fair Information Principles (FIPs) provide a framework to balance privacy and other interests. Below we give an introduction to the FIPs, and to the OECD Privacy Guidelines, which include a version of the FIPs. We also discuss the main challenges when reconciling the FIPs and open data policy.

##### A. FAIR INFORMATION PRINCIPLES (FIPs)

###### 1. *Background of the FIPs*

The Fair Information Principles (FIPs),<sup>146</sup> or the Fair Information Practice Principles (FIPPs),<sup>147</sup> are ingrained in most data privacy laws and guidelines around the world. For example, the FIPs can be recognized in the 1973 report *Records, Computers, and the Rights of Citizens*, by the U.S. Department of Health, Education, and Welfare;<sup>148</sup> the Privacy Act;<sup>149</sup> and the Fair Credit Reporting Act.<sup>150</sup> The Federal Trade Commission and the White House have recently called for FIPs-based privacy regulation for the private sector.<sup>151</sup>

---

146. See NEIL RICHARDS, *INTELLECTUAL PRIVACY: RETHINKING CIVIL LIBERTIES IN THE DIGITAL AGE* 162 (2014) See also Robert Gellman, *Fair Information Practices: A Basic History, Version 2.02*, BOBGELLMAN.COM (2013), <http://bobgellman.com/rg-docs/rg-FIPShistory.pdf>.

147. See *The Fair Information Principles at Work*, U.S. DEP'T OF HOMELAND SECURITY, [http://www.dhs.gov/xlibrary/assets/privacy/dhsprivacy\\_fippsfactsheet.pdf](http://www.dhs.gov/xlibrary/assets/privacy/dhsprivacy_fippsfactsheet.pdf).

148. U.S. DEP'T OF HEALTH, EDUC. & WELFARE, *RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS*, at i, xx (1973), <http://www.justice.gov/opcl/docs/rec-com-rights.pdf>.

149. Privacy Act of 1974, Pub. L. No. 93-579, 88 Stat. 1896 (codified at 5 U.S.C. § 552a (2012)).

150. Fair Credit Reporting Act of 1970, Pub. L. No. 91-508, 84 Stat. 1128 (codified as amended at 15 U.S.C. §§ 1681-1681x (2012)).

151. OFFICE OF THE PRESIDENT, *CONSUMER DATA PRIVACY IN A NETWORKED WORLD* (2012), <http://www.whitehouse.gov/sites/default/files/privacy-final.pdf>; FEDERAL TRADE COMMISSION, *PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS* (2012), <http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf>.

About a hundred countries in the world have a data privacy law including a version of the FIPs.<sup>152</sup> The FIPs can also be recognized in the United Nations Guidelines for the Regulation of Computerized Personal Data Files 1990,<sup>153</sup> and the APEC Privacy Framework of the Asia-Pacific Economic Cooperation (2005).<sup>154</sup> The E.U. Data Protection Directive (1995) contains one of the world's most stringent implementations of the FIPs.<sup>155</sup> European legal scholars tend to speak of data protection principles rather than of FIPs, but both sets of principles are similar.<sup>156</sup> Different countries, however, implement the FIPs differently. The FIPs give guidelines to balance privacy-related interests and other interests, such as those of business and the public sector.<sup>157</sup>

## 2. OECD Guidelines

An influential version of the FIPs can be found in the *Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, from the Organisation for Economic Co-operation and Development (OECD).<sup>158</sup> The OECD was established in 1960, by eighteen European countries, the United States, and Canada.<sup>159</sup> Now, the OECD has thirty-

152. Graham Greenleaf, *Sheherazade and the 101 Data Privacy Laws: Origins, Significance and Global Trajectories*, 23 J.L. INFO. & SCI. 4 (2014); GRAHAM GREENLEAF, GLOBAL TABLES OF DATA PRIVACY LAWS AND BILLS (3d ed. 2013), <http://ssrn.com/abstract=2280875>.

153. Guidelines for the Regulation of Computerized Personal Data Files, G.A. RES. 45/95, U.N. DOC. A/RES/45/95 (Dec. 14, 1990).

154. Asia-Pacific Economic Cooperation [APEC], *Privacy Framework*, APEC Doc. No. 205-SO-01.2 (2005), [http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~media/Files/Groups/ECSG/05\\_ecsg\\_privacyframewk.ashx](http://www.apec.org/Groups/Committee-on-Trade-and-Investment/~media/Files/Groups/ECSG/05_ecsg_privacyframewk.ashx).

155. Directive 95/46/EC, *supra* note 69.

156. The core of E.U. data protection law can be found in article 6 of the Data Protection Directive. Directive 95/46/EC, *supra* note 69, art. 6.

157. Paul de Hert & Serge Gutwirth, *Privacy, Data Protection and Law Enforcement: Opacity of the Individual and Transparency of Power*, in PRIVACY AND THE CRIMINAL LAW 91 (Erik Claes, Antony Duff, & Serge Gutwirth eds., 2006); *see also* RICHARDS, *supra* note 146, at 162; Marc Rotenberg, *Fair Information Practices and the Architecture of Privacy (What Larry Doesn't Get)*, 2001 STAN. TECH. L. REV. 1, 1–4; Ann Cavoukian, *Evolving FIPs: Proactive Approaches to Privacy, Not Privacy Paternalism*, in REFORMING EUROPEAN DATA PROTECTION LAW 293 (Serge Gutwirth, Ronald Leenes & Paul de Hert eds., 2015).

158. *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*, OECD, <http://www.oecd.org/sti/ieconomy/oecdguidelinesontheProtectionofPrivacyandTransborderFlowsOfPersonalData.htm> (last visited June 22, 2015) [hereinafter OECD Privacy Guidelines]. The OECD Privacy Guidelines call these principles “Basic Principles of National Application.”

159. Robert Wolfe, *From Reconstructing Europe to Constructing Globalization: The OECD in Historical Perspective*, in THE OECD AND TRANSNATIONAL GOVERNANCE 25, 25–26 (Rianne Mahon & Stephen McBride eds., 2008).

four member countries, including Mexico, Chile, Korea and Japan.<sup>160</sup> The OECD's self-stated mission is "to promote policies that will improve the economic and social well-being of people around the world."<sup>161</sup>

One of the main reasons for the OECD to adopt the Guidelines was that several European data privacy laws from the 1970s restricted the export of personal data to countries that offered inadequate legal protection to personal data. Some, the United States in particular, worried that European countries would use data privacy law as a trade barrier.<sup>162</sup> The chairman of the expert group that wrote the 1980 OECD Guidelines summarized, "the OECD's central concern was . . . that the response of European nations (and European regional institutions) to the challenges of TBDF [transborder data flows] for privacy might potentially erect legal and economic barriers against which it was essential to provide effective exceptions."<sup>163</sup> Therefore, OECD member states negotiated about more international cooperation, leading to the adoption of the Privacy Guidelines in 1980.<sup>164</sup>

The OECD Guidelines have a dual goal: they aim to protect privacy and individual liberties, and to foster the free flow of information between OECD member countries.<sup>165</sup> Many legal data privacy instruments have a similar dual goal.<sup>166</sup> In this Article we focus on protecting privacy and individual liberties, rather than on transborder data flows.<sup>167</sup>

The Guidelines are not legally binding, they merely "recommend" that OECD member countries implement the Guidelines.<sup>168</sup> The Guidelines

160. *Members and Partners*, OECD, <http://www.oecd.org/about/membersandpartners> (last visited June 22, 2015).

161. *About the OECD*, OECD, <http://www.oecd.org/about> (last visited June 22, 2015).

162. Nicholas Platten, *Background to and History of the Directive*, in EC DATA PROTECTION DIRECTIVE 15 (David Bainbridge ed., 1996); GONZÁLEZ FUSTER, *supra* note 132, at 77.

163. Michael Kirby, *The History, Achievement and Future of the 1980 OECD Guidelines on Privacy*, 20 J.L. INFO. & SCI. 1, 6 (2010).

164. *Id.* at 7–10.

165. OECD Privacy Guidelines, *supra* note 158.

166. See GONZÁLEZ FUSTER, *supra* note 132, at 130. For instance, the E.U. Data Protection Directive has a similar dual goal. See Directive 95/46/EC, *supra* note 69, art. 1.

167. On transborder data flows, see CHRISTOPHER KUNER, TRANSBORDER DATA FLOWS AND DATA PRIVACY LAW (2013).

168. OECD, *Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, at 11–12, C(80)58/FINAL (2013), <http://www.oecd.org/sti/economy/2013-oecd-privacy-guidelines.pdf> (as amended on July 11, 2013).

stress that they provide “minimum standards”<sup>169</sup> and do not “preven[t] the application of different protective measures to different categories of personal data, depending upon their nature and the context in which they are collected, stored, processed or disseminated.”<sup>170</sup> The OECD Guidelines use flexible terms so that all of the member countries can agree with them, even though the United States and European countries have different legal traditions, especially regarding privacy and personal data.<sup>171</sup>

When the OECD Guidelines were adopted in 1980, only about one third of the member states had adopted a data privacy law. Now, almost every OECD member state has a data privacy law with the FIPs at its core.<sup>172</sup> The OECD Guidelines were updated in 2013, but the essence of the principles was retained.<sup>173</sup> The 2013 OECD Privacy Guidelines are listed below. The principles partly overlap, and should be read together:

#### Collection Limitation Principle

There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.<sup>174</sup>

#### Data Quality Principle

Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.<sup>175</sup>

#### Purpose Specification Principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.<sup>176</sup>

---

169. *Id.* at 14.

170. *Id.* at 13.

171. Kirby, *supra* note 163, at 10.

172. David Wright, Paul de Hert, & Serge Gutwirth, *Are the OECD Guidelines at 30 Showing Their Age?*, 54 COMMUNICATIONS OF THE ACM 119, 122 (2011).

173. OECD, THE OECD PRIVACY FRAMEWORK 4, [http://www.oecd.org/sti/ieconomy/oecd\\_privacy\\_framework.pdf](http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf) (last visited June 22, 2015) (“[T]his revision leaves intact the original ‘Basic Principles’ in Part Two of the Guidelines.”).

174. *Id.* at 14 (paragraph 7 of the Guidelines governs the protection of privacy and transborder flows of personal data).

175. *Id.*

176. *Id.*

**Use Limitation Principle**

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with [the Purpose Specification Principle] except:

- a) with the consent of the data subject; or
- b) by the authority of law.<sup>177</sup>

**Security Safeguards Principle**

Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.<sup>178</sup>

**Openness Principle**

There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.<sup>179</sup>

**Individual Participation Principle**

Individuals should have the right:

- a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to them;
- b) to have communicated to them, data relating to them
  - (i) within a reasonable time;
  - (ii) at a charge, if any, that is not excessive;
  - (iii) in a reasonable manner; and
  - (iv) in a form that is readily intelligible to them;
- c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and
- d) to challenge data relating to them and, if the challenge is successful to have the data erased, rectified, completed or amended.<sup>180</sup>

---

177. *Id.*

178. *Id.* at 15.

179. *Id.*

180. *Id.*

### Accountability Principle

A data controller should be accountable for complying with measures which give effect to the principles stated above.<sup>181</sup>

### 3. *Scope of the OECD Guidelines*

The OECD Guidelines apply to “personal data,” which the Guidelines define as “any information relating to an identified or identifiable individual (data subject).”<sup>182</sup> But the Guidelines limit the scope of application considerably; they “apply to personal data, whether in the public or private sectors, which, *because of the manner in which they are processed, or because of their nature or the context in which they are used, pose a risk to privacy and individual liberties.*”<sup>183</sup>

The Guidelines thus follow a risk-based approach: they only apply to personal data processing if it threatens privacy or individual liberties. By contrast, E.U. data protection law generally applies to personal data processing, and requires that personal data be processed fairly, including when the data do not pose a *prima facie* risk for individual liberties.<sup>184</sup>

In this Article, we assume that personal data should always be handled in line with the FIPs.<sup>185</sup> Hence, we do not follow the risk-based approach suggested by the OECD Guidelines. We do consider the risk of personal data processing and the sensitivity of personal data, but we do so *within* the FIPs framework (*see infra* Parts V–VII).

The OECD Guidelines have been criticized, for instance, for implementing the FIPs too weakly. Roger Clarke says the OECD Guidelines aim “to facilitate international business, *not* to protect privacy.”<sup>186</sup> The OECD Guidelines “were motivated by the facilitation of

181. *Id.*

182. *Id.* at 13. The OECD personal data definition is similar to the definition in E.U. data protection law. Directive 95/46/EC, *supra* note 69, art. 2(a).

183. OECD PRIVACY FRAMEWORK, *supra* note 173, at 14 (emphasis added).

184. *See* E.U. Charter of Fundamental Rights, *supra* note 68, art. 8 (“1. Everyone has the right to the protection of personal data concerning him or her. 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law.”); *see also* Case C-131/12, *Google Inc. v Agencia Española de Protección de Datos (AEPD)*, 2014 EUR-Lex CELEX LEXIS 0131 ¶ 69 (May 13, 2014) (CJEU); *Joined Cases 293 & 594/12, Digital Rights Ireland Ltd v Minister for Communications, Marine and Natural Resources, Kärntner Landesregierung*, 2014 EUR-Lex CELEX LEXIS 0293 ¶ 36 (Apr. 8, 2014) (CJEU).

185. The idea that personal data should always be processed in line with the FIPs could be seen as a European approach.

186. Roger Clarke, *Research Use of Personal Data*, Comments at the National Scholarly Communications Forum on Privacy: Balancing the Needs of Researchers and

international business; they were constrained by the need to leave existing legislation unaffected; and their formulation reflected the need for cross-cultural comprehensibility.”<sup>187</sup>

For better or for worse, the FIPs are widely accepted as a starting point for data privacy law. Although the application of the FIPs varies considerably, they express a nearly worldwide consensus on minimum standards for fair personal data use. The next section describes the main challenges that arise when trying to reconcile the FIPs and open data policy.

## B. FIPs AND OPEN DATA: CHALLENGES

To date, policymakers and academics have given limited attention to the question of how privacy norms might be reconciled with policies aimed at making government data available for a wide range of uses. Policymakers and civil society actors recognize the privacy implications of open data.<sup>188</sup> But detailed analyses of the tension between open data and privacy, and especially of open data and the FIPs, is scarce. In the empirical mapping study, we found that, while there were some mentions of open data and privacy together on various forms of digital media, many of these were fleeting or incidental, and few of them contained substantive discussion about how to achieve a balance between the two.<sup>189</sup>

Several open data guidelines from civil society mention privacy—albeit cursorily. For example, the *8 Principles of Open Government Data* state that “[r]easonable privacy, security and privilege restrictions may be allowed.”<sup>190</sup> The Sunlight Foundation says that for a dataset to be open, “[a]ll raw information from [the] dataset should be released to the public, except to

---

the Individual’s Right to Privacy under the New Privacy Laws (Aug. 9, 2002), <http://www.rogerclarke.com/DV/NSCF02.html>; see also William Bonner & Mike Chiasson, *If Fair Information Principles Are the Answer, What Was the Question? An Actor-Network Theory Investigation of the Modern Constitution of Privacy*, 15 INFO. & ORG. 267, 284 (2005).

187. Roger Clarke, *Beyond the OECD Guidelines: Privacy Protection for the 21st Century* (Jan. 4, 2000), <http://www.rogerclarke.com/DV/PP21C.html>.

188. For example, the United Kingdom’s Open Rights Group expressed concern over the U.K. government’s plans to release anonymized health and education data. See *Open Data Privacy*, OPEN RIGHTS GROUP, <https://www.openrightsgroup.org/campaigns/pendata/open-data-privacy> (last visited June 22, 2015). The Open Knowledge and the Open Rights Group convened a working group on open data, personal data, and privacy. See PERSONAL DATA & PRIVACY WORKING GROUP, <http://personal-data.okfn.org/> (last visited June 22, 2015).

189. MAPPING THE POLITICS OF OPEN DATA, *supra* note 62.

190. *8 Principles of Open Government Data*, PUBLIC.RESOURCE.ORG (Dec. 8, 2007), [https://public.resource.org/8\\_principles.html](https://public.resource.org/8_principles.html).

the extent necessary to comply with federal law regarding the release of personally identifiable information.”<sup>191</sup>

In the open data context, governmental and intergovernmental bodies also mention protecting privacy, albeit in a cursory fashion. For example, the *G8 Open Data Charter* recognizes that “there is national and international legislation, in particular pertaining to intellectual property, personally-identifiable and sensitive information, which must be observed.”<sup>192</sup> The 2008 OECD recommendation on public sector information urges that member countries should clearly define “grounds of refusal or limitations,” including “personal privacy.”<sup>193</sup>

Compared to the OECD’s recommendation, the implementation guidance material for Obama’s 2013 executive order contains a more substantive discussion of privacy and the Fair Information Principles. The memorandum includes the suggestion to “[s]trengthen measures to ensure that privacy and confidentiality are fully protected and that data are properly secured[,]” and to “incorporate privacy analyses into each stage of the information’s life cycle.”<sup>194</sup> As well as demanding compliance with relevant laws such as the U.S. Privacy Act of 1974 and the E-Government Act of 2002, the memorandum suggests that “agencies should implement information policies based upon Fair Information Practice Principles and NIST guidance on Security and Privacy Controls for Federal Information Systems and Organizations.”<sup>195</sup>

In the European Union, some work has been done on reconciling privacy and open data, in a thematic network funded by the European Commission to reflect on Legal Aspects of Public Sector Information (LAPSI). The LAPSI Working Group on privacy warns that full application of European data privacy rules will seriously hamper the ability of public sector bodies to disclose information for re-use purposes.<sup>196</sup> In the following section, we discuss the main challenges that occur when trying to reconcile the FIPs and open data policy, starting with the purpose specification principle.

---

191. SUNLIGHT FOUNDATION, *supra* note 12.

192. G8 OPEN DATA CHARTER, *supra* note 1.

193. *OECD Recommendation*, *supra* note 19, at 5.

194. OMB MEMORANDUM M-13-13, OPEN DATA POLICY, *supra* note 25, at 9.

195. *Id.*

196. Dos Santos et al., *supra* note 8, at 348–49; *see also* van Eechoud et al., *LAPSI Position Paper*, *supra* note 8.

### 1. Purpose Specification Principle

The main problem that occurs when trying to reconcile the FIPs and open data policy is that open data policy fosters unanticipated re-use and innovation—“serendipitous reuse” as Shadbolt et al. put it.<sup>197</sup> But secondary use of personal data brings privacy risks. In FIPs parlance, using personal information for unforeseen purposes may breach the purpose specification principle.

The purpose specification principle is a cornerstone of many data privacy laws in the world. It follows from the purpose principle that personal data should only be collected for a purpose that is specified in advance, and that those data should not be used for incompatible purposes.<sup>198</sup> The 1973 “Records, Computers, and the Rights of Citizens” report from the U.S. Department of Health, Education, and Welfare already contained a similar principle: “[t]here must be a way for an individual to prevent information about him that was obtained for one purpose from being used or made available for other purposes without his consent.”<sup>199</sup> In the Charter of Fundamental Rights of the European Union, the purpose specification is included in the right to protection of personal data.<sup>200</sup>

The requirement that personal data may only be used for purposes that are “not incompatible” is somewhat vague. The Article 29 Working Party, an advisory body in which national data protection authorities from Europe cooperate,<sup>201</sup> has discussed the purpose specification in depth. To

197. Nigel Shadbolt, Wendy Hall & Tim Berners-Lee, *The Semantic Web Revisited*, 21 IEEE INTELLIGENT SYSTEMS 96, 98 (2006); see also WENDY HALL ET AL., NOMINET TRUST, OPEN DATA AND CHARITIES 16 (2012), <http://www.nominettrust.org.uk/sites/default/files/Open%20Data%20and%20Charities.pdf> (“Open data, taking inspiration from other ideologies of openness such as open source and open access publishing, articulates the idea that data should be usable by anyone, not just the data owner (or ‘data controller’ in the language of the Data Protection Act).”).

198. See the Purpose Specification Principle from the OECD Guidelines, *supra* Section IV.A.2.

199. RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS, *supra* note 148, at xx.

200. E.U. Charter of Fundamental Rights, *supra* note 68, art. 8(2).

201. On the Working Party generally, see Yves Poullet & Serge Gutwirth, *The Contribution of the Article 29 Working Party to the Construction of a Harmonised European Data Protection System: An Illustration of “Reflexive Governance”?*, in DÉFIS DU DROIT À LA PROTECTION DE LA VIE PRIVÉE [CHALLENGES OF PRIVACY AND DATA PROTECTION LAW] 570 (María Verónica Perez Asinari & Pablo Palazzi eds., 2008). The Working Party’s opinions are not legally binding, but they are influential in Europe. Judges and national Data Protection Authorities often follow the Working Party’s interpretation.

assess whether a new purpose is compatible with the collection purpose, says the Working Party, all circumstances must be considered. Relevant circumstances include the relation between the original and the new purpose, the collection context, the reasonable expectations of the data subject,<sup>202</sup> the personal data's sensitivity, the risks resulting from the new purpose, and the measures the controller has in place to mitigate risks.<sup>203</sup>

According to the Working Party, an example of a new purpose that is incompatible with the original processing purpose is contained in the following hypothetical. A public sector body publishes public servants' contact details on its website, to enable the public to contact them.<sup>204</sup> A re-user wants to merge the public servants' home addresses and phone numbers with the published contact details, to build an interactive map.<sup>205</sup> The re-use is not within the reasonable expectations of the civil servants, making the purpose incompatible and thus not allowed.<sup>206</sup>

## 2. *Security and Accountability Principles*

The security principle requires appropriate security for personal data. Data controllers must protect data against unauthorized disclosure, access, or other use. When thoughtlessly releasing personal data, a public sector body breaches the security principle. After all, the public sector body would have no control over how the data are used—and neither would the data subjects. The mere fact that data subjects have no control over the use of their data is a subjective privacy harm. Moreover, anybody could access the data, including data brokers and identity thieves.

The accountability principle makes the data controller responsible for complying with the FIPs. The OECD Guidelines define the data controller as the party that “is competent to decide about the contents and use of personal data regardless of whether or not such data are collected, stored, processed or disseminated by that party or by an agent on its

---

202. In the United States and the European Union, the “reasonable expectation of privacy” is interpreted differently. The European Court of Human Rights says a “person’s reasonable expectations as to privacy is a significant though not necessarily conclusive factor.” *Perry v. United Kingdom*, 2003-IX Eur. Ct. H.R. 141, ¶ 37. On the United States, see SOLOVE & SCHWARTZ, *supra* note 66, at 288–335.

203. *Opinion of the Article 29 Data Protection Working Party on Open Data and Public Sector Information (“PSI”) Re-use*, at 20, 1021/00/EN WP 207 (June 5, 2013) [hereinafter *Article 29 Opinion on Open Data and PSI Re-use*]; see also *Opinion of the Article 29 Data Protection Working Party on Purpose Limitation*, 00569/13/EN WP 203 (Apr. 2, 2013).

204. *Article 29 Opinion on Open Data and PSI Re-use*, *supra* note 203, at 20.

205. *Id.*

206. *Id.*

behalf.”<sup>207</sup> A public sector body holding the personal data is usually the data controller. If a re-user obtains personal data from the public sector body, the re-user typically becomes a data controller as well.

### 3. *Data Quality Principle*

The data quality principle requires appropriate accuracy, completeness, and relevancy of personal data. One of the aims of the principle is to reduce the risk that organizations base decisions about people on incorrect data. Decisions based on incorrect data can have disastrous effects for a data subject.<sup>208</sup> The data quality principle is relevant to open data. Releasing incorrect personal data could have a detrimental effect.<sup>209</sup> For example, imagine that a website about political campaign financing erroneously includes your name as a donor to a fringe extremist party.

### 4. *Collection Limitation and Transparency Principle*

The transparency principle, or openness principle, requires transparency regarding data processing, especially towards the data subject.<sup>210</sup> The transparency principle aims to prevent data controllers from abusing information asymmetry.

The transparency principle is prominent in data privacy laws, and can be recognized, for instance, in the proposed U.S. Consumer Privacy Bill of Rights,<sup>211</sup> the E.U. Data Protection Directive,<sup>212</sup> and the proposed E.U. Data Protection Regulation.<sup>213</sup> Some authors suggest that the transparency

207. OECD Privacy Guidelines, *supra* note 158, art. 1. The OECD data controller concept is different from the E.U. concept of “data controller.” In brief, under E.U. data protection law, the data controller is the party that determines the goals and means for personal data processing. A party that processes personal data on behalf of the controller is the “data processor.” Directive 95/46/EC, *supra* note 69, art. 2(d)–(e).

208. *See, e.g.,* Romet v. Netherlands, Eur. Ct. H.R. No. 7094/06 (2012).

209. *See* Scassa, *supra* note 71; Rotenberg, *supra* note 157.

210. To avoid confusion with the open character of open data, we will speak of the “transparency principle” rather than of the “openness principle.”

211. CONSUMER DATA PRIVACY IN A NETWORKED WORLD, *supra* note 151, at 47 (discussing the Consumer Privacy Bill of Rights and transparency principle).

212. Directive 95/46/EC, *supra* note 69, arts. 10, 11.

213. *Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, art. 5(a), COM (2012) 11 final (Jan. 25, 2012). In December 2015, agreement was reached on the Regulation’s text. *See Regulation (EU) No. XXX/2016 of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation)*, COM (2016) 15039/15 limite (Dec. 15, 2015), <http://data.consilium.europa.eu/doc/document/ST-15039-2015-INIT/en/pdf>. At the time of writing, the European Parliament and the Council must

principle is the most important principle of the FIPs.<sup>214</sup> The transparency principle has old roots. The first principle of the U.S. Department of Health, Education, and Welfare report of 1973 states “[t]here must be no personal-data record-keeping systems whose very existence is secret.”<sup>215</sup> The second principle adds that “[t]here must be a way for an individual to find out what information about him is in a record and how it is used.”<sup>216</sup>

The collection limitation principle requires that personal data, where appropriate, be collected with the data subject’s knowledge or consent. The Article 29 Working Party recommends that a public sector body inform data subjects in advance whether the personal data they provide might be disclosed, for example due to freedom of information laws.<sup>217</sup>

### 5. *Use Limitation and Individual Participation Principle*

The individual participation principle aims to give people some control over the processing of their personal data. For instance, data subjects have the right, under certain circumstances, to rectify their data. The principle illustrates that the privacy as control perspective has influenced the FIPs.<sup>218</sup>

As previously stated, unrestricted re-use of personal data would breach the purpose specification principle—but the use limitation principle seems to offer a way out. The use limitation principle says that personal data should only be used in accordance with the purpose specification principle, except “a) with the consent of the data subject; or b) by the authority of law.”<sup>219</sup> Hence, personal data can be used for a new (*prima facie* incompatible) purpose if the data subject consents to the new use. Indeed, some have suggested that public sector bodies should obtain consent of the relevant individuals before releasing data as open data.<sup>220</sup>

---

still formally adopt the final text. The Regulation is expected to become applicable in 2018.

214. *See, e.g.,* de Hert & Gutwirth, *supra* note 157; ZUIDERVEEN BORGESIUS, IMPROVING PRIVACY PROTECTION, *supra* note 73, at 99, 106–11.

215. RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS, *supra* note 148, at 41.

216. *Id.*

217. *Article 29 Opinion on Open Data and PSI Re-use, supra* note 203, at 9. Arvind Narayanan et al. suggest that people should be informed regarding re-identification risks. ARVIND NARAYANAN, JOANNA HUEY & EDWARD W. FELTEN, A PRECAUTIONARY APPROACH TO BIG DATA PRIVACY 1, 16 (2015).

218. *See* Kirby, *supra* note 163, at 8 (citing Alan Westin as an influence on the OECD Guidelines).

219. OECD PRIVACY FRAMEWORK, *supra* note 173, ¶ 10 (Use Limitation Principle).

220. *See, e.g.,* Bart van der Sloot, *On the Fabrication of Sausages, or of Open Government and Private Data* 3 JeDEM 1, 14 (2011).

However, relying on data subject consent for disclosing personal data as open data has some drawbacks. First, people are often in a dependent position vis-à-vis the public sector, and that position may make consent involuntary. Somebody interacting with the public sector might not feel free to withhold consent. Say Alice goes to a city council office for unemployment benefits. Alice really needs money, as she has missed five rent payments, and risks being evicted with her young child. Because she wants to be cooperative, Alice is unlikely to withhold consent to any request by the city council office. Under E.U. data privacy law, consent given under too much pressure is invalid, because consent must be “freely given.”<sup>221</sup> For instance, if an employer asks an employee for consent, the consent might not be freely given because of the power imbalance.<sup>222</sup> And according to the European Court of Justice, people applying for passports cannot be deemed to have freely consented to have their fingerprints taken, because people need a passport.<sup>223</sup>

A second problem with data subject consent as a justification for disclosing personal data is that a request for consent can only be meaningful if it specifies a processing purpose.<sup>224</sup> A third problem is that behavioral studies cast doubt on individual consent as a privacy protection measure. For example, on the Internet, people tend to click “I agree” to requests that they see on their screens without knowing what they are agreeing to.<sup>225</sup> Furthermore, it may be impractical for the public sector body to obtain the consent of thousands of individuals. In sum, obtaining data subjects’ consent to release personal data is not a general solution to reconcile FIPs and open data policy.

To conclude, from a FIPs perspective, the main problem with open data is that it can be used by anyone, for any purpose, without re-use

---

221. ELENI KOSTA, CONSENT IN EUROPEAN DATA PROTECTION LAW 256 (2013).

222. *Opinion of the Article 29 Data Protection Working Party on the Definition of Consent*, at 13–14, 01197/11/EN WP187 (July 13, 2011) [hereinafter *Article 29 Opinion on Consent*].

223. C-291/12, Schwarz v. Stadt Bochum, EUR-Lex CELEX LEXIS 0291 ¶ 32 (Oct. 17, 2013) (CJEU).

224. *Article 29 Opinion on Consent*, *supra* note 222, at 9.

225. See, e.g., Acquisti & Grossklags, *supra* note 93; Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013); Solon Barocas & Helen Nissenbaum, *Big Data’s End Run Around Anonymity and Consent*, in *PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT* 44 (Julia Lane et al. eds., 2014); ZUIDERVEEN BORGESIOUS, *IMPROVING PRIVACY PROTECTION*, *supra* note 73.

restrictions. A complete lack of re-use restrictions clashes with the purpose specification principle.

## V. TYPES OF DATA

Compromises are possible to balance privacy and open data interests. This balancing act may play out differently for different types of data. To help balance the different interests, we distinguish between four data categories, with different levels of privacy risks: (A) raw personal data, (B) pseudonymized data, (C) anonymized data, and (D) non-personal data. We borrow the “raw personal data” category from Tim Davies, and borrow the other three categories from the Article 29 Working Party.<sup>226</sup> We distinguish the four categories to structure the discussion, but the boundaries between them are not clear-cut, as noted in Section V.E.

### A. RAW PERSONAL DATA

With raw personal data, no attempt has been made to mitigate re-identification risks. Examples of raw personal data include names, social security numbers, and personal email addresses. Some open data advocates suggest that open data should never include raw personal data.<sup>227</sup> Indeed, while open data policy is important, releasing raw personal data without any re-use restrictions is usually neither desirable nor legally feasible.

However, in some circumstances raw personal data should be disclosed, because the public interests in disclosure outweigh the privacy interests. For instance, say a public registry of judges reveals positions and jobs judges hold elsewhere, to uphold impartiality of the judiciary. If these data did not identify individual judges, disclosure would not offer

---

226. Tim Davies, *Untangling the Data Debate: Definitions and Implications*, OPENDATAIMPACTS.NET (Mar. 23, 2012), <http://www.opendataimpacts.net/2012/03/untangling-the-open-data-debate-definitions-and-implications>; *Opinion of the Article 29 Data Protection Working Party on Anonymisation Techniques*, 0829/14/EN WP 216, (Apr. 10, 2014) [hereinafter *Article 29 Opinion on Anonymisation Techniques*]. See also HALL ET AL., NOMINET TRUST, *supra* note 197.

227. For instance, Wendy Hall et al. say that raw personal data “should never be directly published as openly licensed and accessible data without explicit consent of the individuals covered in the data.” HALL ET AL., NOMINET TRUST, *supra* note 197, at 14. Tim Berners-Lee and Nigel Shadbolt, two authors who promote open data, say that “[i]n the drive to free up data we have always argued that it is essential to respect individual privacy and national security.” Tim Berners-Lee & Nigel Shadbolt, *There’s Gold to be Mined from All Our Data*, TIMES (London), Dec. 31, 2011.

sufficient transparency.<sup>228</sup> More generally, people must accept that their privacy diminishes if they take on certain functions in the public sector. For example, it is widely accepted that media can report on politicians, even when politicians might sometimes prefer that certain information remain confidential.<sup>229</sup>

But the fact that certain raw personal data should be disclosed does not imply that they should be disclosed as open data without re-use restrictions. Even if a law states that certain information must be made public, it does not necessarily follow that such information should be released fully openly. By 1972, some already argued that “the assumptions built into 19th century ideals of public records need revisiting in light of technology.”<sup>230</sup> And as Scassa notes, “[i]n many cases, decisions around the public nature of the information were made in an era before the Internet.”<sup>231</sup> Hence, personal data that are required to be public by law should not automatically be seen as data that can be released as fully open data.

To illustrate, in many countries court proceedings are mandated to be public. But if court proceedings can only be consulted by traveling to the courthouse and inspecting paper files, the personal information in those files is protected by “practical obscurity.”<sup>232</sup> As the U.S. Supreme Court noted in 1989, “there is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information.”<sup>233</sup>

In sum, even if the law requires disclosing certain personal information as part of the public record, the public sector body should still assess whether this information should also be made available as open data on

---

228. As Gary T. Marx puts it, sometimes “disclosure norms” trump “privacy norms.” Gary T. Marx, *Foreword: Privacy Is Not Quite Like the Weather*, in *PRIVACY IMPACT ASSESSMENT* v, viii (David Wright & Paul De Hert eds., 2012).

229. See *supra* Section III.C.

230. Chris Jay Hoofnagle, Summary, *Archive of the Meetings of the Secretary’s Advisory Committee on Automated Personal Data Systems (SACAPDS)*, *BERKELEY LAW* (July 15, 2014), <https://www.law.berkeley.edu/centers/bclt/research/privacy-at-bclt/archive-of-the-meetings-of-the-secretarys-advisory-committee-on-automated-personal-data-systems-sacapds>.

231. Scassa, *supra* note 71, at 403; see also Keenan, *supra* note 97, at 1.

232. *U.S. Dep’t of Justice v. Reporters Comm. for Freedom of the Press*, 489 U.S. 749, 762 (1989).

233. *Id.* at 764.

the web.<sup>234</sup> As we shall argue in Part VI, the question of whether or not data should be made available as open data is a further, additional question that follows the question of whether the data should be made publicly available at all.

#### B. PSEUDONYMIZED DATA

Pseudonymized data are personal data about an individual that are tied to a unique identifier other than a name. For instance, “William Carey Jones” could be referred to as person number “4.417.749.” Pseudonymization can be described as follows: “replacing one attribute (typically a unique attribute) in a record by another.”<sup>235</sup> Merely substituting names with other unique identifiers is rarely enough to anonymize personal data, or to safeguard privacy.<sup>236</sup>

A well-known example of the limited effect of pseudonymization as an anonymization measure is the 2006 AOL data breach. AOL released pseudonymized data about users of its search engine by replacing the name of each searcher with a number.<sup>237</sup> However, journalists soon found out the real name of the person behind one of the pseudonymous search profiles and published an article entitled *A Face is Exposed for AOL Searcher No. 4417749*.<sup>238</sup> The journalists found the woman behind search profile 4417749 without using sophisticated re-identification techniques.<sup>239</sup> The search queries of user number 4417749 suggested that the searcher was an elderly woman with a dog, from a specific town.<sup>240</sup> When the journalists visited her house, she confirmed that the searches were hers.<sup>241</sup>

The Article 29 Working Party suggests that pseudonymized data are a type of personal data, and are thus within the scope of European data protection law.<sup>242</sup> Some computer scientists have a similar view.<sup>243</sup>

---

234. See generally Amanda Conley, Anupam Datta, Helen Nissenbaum & Divya Sharma, *Sustaining Privacy and Open Justice in the Transition to Online Court Records: A Multidisciplinary Inquiry*, 71 MD. L. REV. 772 (2012).

235. *Article 29 Opinion on Anonymisation Techniques*, *supra* note 226, at 20.

236. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 2; PRESIDENT’S COUNCIL OF ADVISORS ON SCI. & TECH., EXEC. OFFICE OF THE PRESIDENT, *BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE*, at 38–39 (2014), [http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_big\\_data\\_and\\_privacy\\_-\\_may\\_2014.pdf](http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf).

237. Michael Barbaro & Tom Zeller Jr., *A Face is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES, Aug. 9, 2006.

238. *Id.*

239. *Id.*

240. *Id.*

241. *Id.*

242. *Article 29 Opinion on Anonymisation Techniques*, *supra* note 226, at 10.

243. NARAYANAN, HUEY & FELTEN, *supra* note 217.

However, the Working Party's view has also been criticized for making the scope of personal data too broad.<sup>244</sup>

While pseudonymizing personal data rarely, if ever, makes people non-identifiable, pseudonymization can help to protect privacy interests, by making it a bit harder to recognize people by name.<sup>245</sup> For instance, Conley et al. suggest that pseudonymization can help to mitigate privacy concerns when court cases are published online.<sup>246</sup> If people's names are changed to "[party 1]" and "[party 2]" in judgments, it would be impossible to search within court records on the basis of a person's name. Pseudonymization also reduces the chance that somebody who looks at the data will recognize a person by name. However, it might still be possible to recognize people based on the facts of a case discussed in the judgment. Nevertheless, pseudonymization adds a thin layer of practical obscurity.<sup>247</sup>

Different countries have different traditions. In the Netherlands, many court decisions are published online on a centralized website.<sup>248</sup> But if the litigating parties are individuals, their names are changed to neutral phrases such as "plaintiff" and "defendant."<sup>249</sup> In other countries, litigants' names are often included in court documents, even when published online.<sup>250</sup>

---

244. See, e.g., Khaled El Emam & Cecilia Álvarez, *A Critical Appraisal of the Article 29 Working Party Opinion 05/2014 on Data Anonymization Techniques*, 5 INT'L DATA PRIVACY L. 73 (2015).

245. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 2; PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., *supra* note 236, at 38–39.

246. Conley, Datta, Nissenbaum & Sharma, *supra* note 234, at 842.

247. See *id.*

248. See generally Laurens Mommers, *Access to Law in Europe*, in INNOVATING GOVERNMENT 383 (Simone van der Hof & Marga M. Groothuis eds., 2011); LEONIE VAN LENT, EXTERNE OPENBAARHEID IN HET STRAFPROCES (2008).

249. These are our translations. *Anonimiseringsrichtlijnen*, RECHTSpraak, <https://www.rechtspraak.nl/Uitspraken-en-nieuws/Uitspraken/Paginas/Anonimiseringsrichtlijnen.aspx> (last visited July 2, 2015). Not all personal data are obfuscated; for example, the attorneys for a case are mentioned by name. The Spanish system is similar to the Dutch one. See James B. Jacobs & Elena Laurrauri, *Are Criminal Convictions a Public Matter? The USA and Spain*, 14 PUNISHMENT & SOC'Y 3 (2012) (with further references to literature on other countries).

250. About the United States, see Nancy S. Marder, *From "Practical Obscurity" to Web Disclosure: A New Understanding of Public Information*, 59 SYRACUSE L. REV. 441, 444–47 (2009); Conley, Datta, Nissenbaum & Sharma, *supra* note 234. For an overview of how to obtain criminal records in fifty-four countries, see KPMG, DISCLOSURE OF CRIMINAL RECORDS IN OVERSEAS JURISDICTIONS (2009), [http://www.cpni.gov.uk/documents/publications/2009/2009-criminal\\_records\\_disclosure\\_intro\\_and\\_exe\\_summary\\_march09.pdf?epslanguage=en-gb](http://www.cpni.gov.uk/documents/publications/2009/2009-criminal_records_disclosure_intro_and_exe_summary_march09.pdf?epslanguage=en-gb).

In sum, pseudonymization can help to reduce privacy risks—it is a useful but not sufficient security measure. Because pseudonymizing data is not enough to anonymize data, pseudonymous data must generally be treated as personal data.

### C. ANONYMIZED DATA

Anonymized data are ex-personal data that are rendered anonymous in such a way that data subjects are no longer identifiable. Aggregated data are typically anonymous. For instance, the information that “112,580 people live in Berkeley,” without additional information, does not identify an individual. Anonymization can be defined as “a technique applied to personal data in order to achieve irreversible de-identification.”<sup>251</sup> Anonymized data are outside the scope of the FIPs, as the FIPs only apply to personal data.

The fact that personal data can be aggregated and thereby anonymized seems an appropriate way to strike a balance between privacy interests and open data interests.<sup>252</sup> For instance, statistics can often be disclosed as open data, as long as they are anonymized and aggregated.<sup>253</sup> To illustrate, a crime map could say that on a certain day, “between one and ten burglaries took place on or near Bancroft Way,” rather than “one burglary took place at 11 Bancroft Way.” Traffic data could say that “between one and ten cars drove on Bancroft Way between April 15 and April 20, 2015,” rather than “one car drove on Bancroft Way on April 19, 2015 at 12:24 A.M.”

But two caveats are in order. First, anonymizing data does not guarantee privacy and fairness.<sup>254</sup> For instance, the Dutch public reacted angrily when the police used aggregated information derived from data gathered by TomTom, a vendor of car navigation systems.<sup>255</sup> The police

---

251. *Article 29 Data Protection Working Party on Anonymisation Techniques*, *supra* note 226, at 7.

252. *See Article 29 Opinion on Open Data and PSI Re-use*, *supra* note 203, at 12; HALL ET AL., NOMINET TRUST, *supra* note 197, at 45; *see also* OFFICE OF THE AUSTRALIAN INFO COMM’R, INFORMATION POLICY AGENCY RESOURCE 1: DE-IDENTIFICATION OF DATA AND INFORMATION, (2014), [http://www.oaic.gov.au/images/documents/information-policy/information-policy-resources/information-policy-agency-resources/information\\_policy\\_agency\\_resource\\_1.pdf](http://www.oaic.gov.au/images/documents/information-policy/information-policy-resources/information-policy-agency-resources/information_policy_agency_resource_1.pdf).

253. *See* Francesco Molinari & Jesse Marsh, *Does Privacy Have to Do with Open Data? Some Preliminary Reflections—and Answers*, in CEDEM13 CONFERENCE FOR E-DEMOCRACY AND OPEN GOVERNMENT 303, 311 (Peter Parycek & Noella Edelman eds., 2013).

254. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 3.

255. Charles Arthur, *TomTom Satnav Data Used to Set Police Speed Traps*, *GUARDIAN*, Apr. 28, 2011.

used the data to choose the best spots to install speeding cameras.<sup>256</sup> The Dutch Data Protection Authority examined whether TomTom's practices complied with E.U. data privacy law, and did not find major problems.<sup>257</sup> The data obtained by the police were properly anonymized through aggregation, and thus outside the scope of the FIPs.<sup>258</sup>

The TomTom example illustrates a broader problem: the FIPs apply to personal data—and only to personal data. But people can be treated unfairly, or feel like they are being treated unfairly, on the basis of information that is *based* on personal data concerning them, but that is not personal data anymore.<sup>259</sup> Moreover, as the aggregated information is outside the scope of the FIPs, the data subject rights that follow from the FIPs, such as access and correction rights, no longer apply. As Seda Gürses notes, anonymization can “disempower” the individual.<sup>260</sup> The FIPs and most data privacy laws around the world have this problem in common.<sup>261</sup> We will not attempt to solve the problem here. But we do note that sometimes a public sector body may want to decide not to release anonymized information, even if the information is outside the of FIPs' scope.

A second caveat is that anonymized data are often less interesting for re-users than raw personal data or pseudonymous data. As Bendert Zevenbergen et al. put it:

The utility and privacy of data are generally directly and inversely related. For many datasets, it has proven difficult—if not impossible—to increase data subjects' privacy without concurrently decreasing the overall utility of the dataset. Small privacy gains are generally achieved by far-reaching decreases in

---

256. *Id.*

257. Press Release, *Following Report by Dutch DPA, TomTom Provides User with Better Information*, COLLEGE BESCHERMING PERSOONSGEGEVENS (Jan. 12, 2012), <https://cbpweb.nl/en/news/following-report-dutch-dpa-tomtom-provides-user-better-information>.

258. *See id.*; *see also* Harold Goddijn, *This is What We Really do with Your Data*, TOMTOM.COM, <http://www.tomtom.com/page/facts> (last visited June 23, 2015).

259. *See generally* Lyon, *Surveillance as Social Sorting*, *supra* note 102.

260. Seda Gürses, *The Spectre of Anonymity*, in SNIFF, SCRAPE, CRAWL . . . : ON PRIVACY, SURVEILLANCE AND OUR SHADOWY DATA-DOUBLE 52 (2012).

261. *See generally* PROFILING THE EUROPEAN CITIZEN: CROSS-DISCIPLINARY PERSPECTIVES (Mireille Hildebrandt & Serge Gutwirth eds., 2008); Barocas & Nissenbaum, *supra* note 225; Joris Van Hoboken & Frederik Zuiderveen Borgesius, *Scoping Electronic Communication Privacy Rules: Data, Services or Values* (2015).

data utility. A small increase in data utility often requires much more personal information to be revealed.<sup>262</sup>

In sum, anonymized data can—in theory—safely be disclosed as open data, without re-use restrictions. However, in practice, irreversible anonymization is exceedingly difficult, and perhaps impossible.

#### D. NON-PERSONAL DATA

A fourth type of data is non-personal data. Many datasets do not contain, and have never contained, personal data. Examples include datasets regarding public transport times, weather conditions, sea tides, road maps, public sector budgets, and environmental pollution.<sup>263</sup> Such datasets have little to do with information about individuals, and do not fall under the purview of the FIPs.

The FIPs do not hinder releasing datasets with non-personal data. Hence, strict compliance with the FIPs does not necessarily interfere with releasing public sector information. Some suggest that “[m]ost open datasets have nothing personal to be protected in them (e.g.: digital maps, public budgets, air pollution measurements etc.)”<sup>264</sup>

But even for non-personal data, there are caveats. First, sometimes there may be non-privacy related arguments against releasing data. For instance, some information may have to remain confidential because of state security, such as information regarding critical infrastructure locations.<sup>265</sup> Second, as discussed in the next section, a dataset with non-personal data may, on closer inspection, include information about an individual.

In sum, we distinguish between four data categories with different risk levels: raw personal data, pseudonymous data, anonymized data, and non-personal data. However, the categories cannot be neatly distinguished in practice, as discussed next.

---

262. BENDERT ZEVENBERGEN, IAN BROWN, JOSS WRIGHT & DAVID ERDOS, OXFORD INTERNET INST., *ETHICAL PRIVACY GUIDELINES FOR MOBILE CONNECTIVITY MEASUREMENTS* 11 (2013), [http://www.oii.ox.ac.uk/research/Ethical\\_Privacy\\_Guidelines\\_for\\_Mobile\\_Connectivity\\_Measurements.pdf](http://www.oii.ox.ac.uk/research/Ethical_Privacy_Guidelines_for_Mobile_Connectivity_Measurements.pdf). Along similar lines, see Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 *UCLA L. REV.* 1701 (2010). Slightly more optimistic is Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 *U. COLO. L. REV.* 1117 (2013).

263. Molinari & Marsh, *supra* note 253, at 311.

264. *Id.*; see also NARAYANAN, HUEY & FELTEN, *supra* note 217, at 21.

265. See Conley, Datta, Nissenbaum & Sharma, *supra* note 234, at 827.

## E. FUZZY BOUNDARIES

The borders between the four data categories are fuzzy. While many data privacy laws make a distinction between personal data and anonymized data, computer science suggests that the distinction is a matter of degree rather than kind.<sup>266</sup> Irreversible anonymization is difficult—perhaps impossible.<sup>267</sup> Apart from that, it is possible to distinguish sub-categories within the four categories. For instance, Zevenbergen et al. distinguish between three types of purportedly anonymized data, with different levels of re-identification risk.<sup>268</sup> And it is debatable whether data about an individual tied to his or her social security number should be seen as raw personal data or as pseudonymized data.

Whether data are sufficiently anonymized is difficult to assess in advance. This is especially so, as more datasets may become available that enable “jigsaw identification.”<sup>269</sup> The more data public sector bodies release, the higher the potential for combining data and thus creating information that can identify people.<sup>270</sup> The Obama administration recognizes this, and urges departments and agencies to perform a risk analysis.<sup>271</sup>

Even purportedly non-personal data can provide information about an individual. For example, a dataset with local air pollution levels contains non-personal data. However, if the dataset says that zip code 94720 is the most polluted, and the only business in that zip code is a one-man business, the pollution level in that zip code can say something about the business owner—namely that he or she is likely polluting. We do not suggest that privacy should enable business owners to escape responsibility for polluting. We merely want to illustrate that even datasets with non-

---

266. See, e.g., Arvind Narayanan & Vitaly Shmatikov, *Myths and Fallacies of “Personally Identifiable Information,”* 53 COMM. ACM 24 (2010); Ohm, *supra* note 262; Matthijs R. Koot, *Measuring and Predicting Anonymity* (2012) (Ph.D thesis, University of Amsterdam), [https://cyberwar.nl/d/PhD-thesis\\_Measuring-and-Predicting-Anonymity\\_2012.pdf](https://cyberwar.nl/d/PhD-thesis_Measuring-and-Predicting-Anonymity_2012.pdf).

267. Narayanan & Shmatikov, *supra* note 266, at 26.

268. ZEVENBERGEN, BROWN, WRIGHT & ERDOS, *supra* note 262, at 22.

269. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 5–7. The phrase “jigsaw identification” is from Kieron O’Hara’s book. O’HARA, *supra* note 74, at 40.

270. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 5–7; see also U.S. GOV’T ACCOUNTABILITY OFFICE, *supra* note 79, at 107.

271. There, the risk is called the “mosaic effect.” OMB MEMORANDUM M-13-13, OPEN DATA POLICY, *supra* note 25, at 9–10.

personal data can provide information about an individual, for instance after linking datasets.<sup>272</sup>

In conclusion, we distinguish between four data categories with different risks levels: raw personal data, pseudonymous data, anonymized data, and non-personal data. The next section shows that open data should not be considered the only route when arguments for disclosure outweigh privacy interests. Options other than releasing data as open data are also available, such as disclosing data with access or re-use restrictions.

## VI. TYPES OF DISCLOSURE

A maximalist approach to publishing public sector information as open data might imply that a public sector body should not impose any conditions on accessing or re-using public sector information. But a more moderate view is that public sector bodies should be allowed to impose conditions for access and re-use, if this is required to protect privacy interests.

We distinguish between three types of disclosure with different degrees of openness: (A) restricted access, (B) restricted use, and (C) open data. Restrictions on access and restrictions on re-use can be combined. Some forms of access and re-use restrictions do not comply with certain definitions of “open” data.<sup>273</sup> But sometimes disclosing data with restrictions is better than not disclosing at all.<sup>274</sup>

### A. DISCLOSURE WITH ACCESS RESTRICTIONS

The first way to balance privacy and open data policy is by restricting access. To achieve a particular objective that underpins open data, it might not be necessary to allow everyone access, or to allow access to the raw data held by a public sector body. Data can be disclosed to particular groups for particular purposes, rather than to anybody for any purpose. Completely blocking data release on the one hand, and releasing data as

---

272. For example, a health insurance company might use the data to calculate the health risks of the pollution, and might charge some people higher prices for coverage.

273. For example, if data are made available for non-commercial uses only, this runs counter to the open data principles set out in Part II (“open” implies that data can be used for any purpose). The same is true if only certain types of users are given access (“open” implies that data can be used by anyone). *See supra* Section II.A.

274. Scassa arrives at a similar conclusion about balancing privacy and public access. “As the experience of courts and tribunals shows, it may sometimes be necessary to place limits on the digital disclosure of some of the information in a ‘public’ record in order to achieve this balance.” Scassa, *supra* note 71, at 404.

open data on the other hand, can be seen as two extremes on a continuum. Disclosing data with restrictions is in between those two extremes.

There are various ways to disclose data, which bring different risk levels. For example, Zevenbergen et al. distinguish open data from “restricted” disclosure, “managed access,” “interactive methods,” and “hybrid” methods.<sup>275</sup>

With “restricted” disclosure, data are only disclosed “to persons or organisations on request, refusing dissemination when the level of risk is considered too high.”<sup>276</sup> Zevenbergen et al. suggest, for instance, that it is riskier to disclose data to a company than to academic researchers.<sup>277</sup> One problem with this type of disclosure is that it is hard to monitor what a receiving party does with the data.

With managed access, “[t]hird parties can query the dataset and conduct statistical (or other) analysis. Such an approach allows the researcher to ascertain exactly who accesses the datasets, while maintaining control over its dissemination.”<sup>278</sup> For instance, researchers might have to visit the offices of the public sector body to inspect data.<sup>279</sup>

An example of an interactive method is “differential privacy.”<sup>280</sup> As Zevenbergen et al. explain:

Differential Privacy . . . only gives statistical answers to queries about an underlying dataset. To protect privacy even further, a certain amount of noise is added to the disclosed statistical data. In principle, differential privacy offers a lower risk for privacy, but there are certain limitations to this approach that need to be understood. For example, the uncertainty related by the addition of noise to the data can be exhausted, which means the dissemination must then stop.<sup>281</sup>

---

275. ZEVENBERGEN, BROWN, WRIGHT & ERDOS, *supra* note 262, at 28–29.

276. *Id.* at 28.

277. *Id.* at 14–16. Similarly, Narayanan et al. say that restricted access “is a good solution” to enable scientific research without releasing data as fully open data. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 20.

278. ZEVENBERGEN, BROWN, WRIGHT & ERDOS, *supra* note 262, at 29.

279. *See id.* at 15.

280. Cynthia Dwork, *Differential Privacy*, in *ENCYCLOPEDIA OF CRYPTOGRAPHY AND SECURITY* 338, 338–40 (Henk C.A. van Tilborg & Sushil Jajodia eds., 2011).

281. ZEVENBERGEN, BROWN, WRIGHT & ERDOS, *supra* note 262, at 29 (emphasis omitted).

Hybrid approaches are also possible. For instance, parts of a dataset could be disclosed publicly, while other parts of the set could be kept confidential, or could be disclosed with strict access restrictions.<sup>282</sup>

In sum, sometimes a compromise between openness and privacy can be found by releasing data with access restrictions. Apart from access restrictions, it is also possible to restrict re-use, as discussed next.

## B. DISCLOSURE WITH RE-USE RESTRICTIONS

Another way to strike a balance between privacy and open data policy is by applying restrictions on re-use of the disclosed data.<sup>283</sup> For instance, re-use restrictions can come in the form of licenses.<sup>284</sup> The license could require re-users not to re-identify data. Such measures have been used in practice. For example, on the website of the U.S. Healthcare Cost and Utilization Project, data users can purchase data sets—but if they purchase a dataset, they must sign an agreement that “expressly prohibits any attempt to identify individuals.”<sup>285</sup>

The Article 29 Working Party suggests that the license should “prohibit license-holders from using the data to take any measure or decision with regard to the individuals concerned.”<sup>286</sup> The license should also require “the license-holder to notify the licensor in case it is detected that individuals can be or have been re-identified.”<sup>287</sup> As proper anonymization is difficult, in some situations, anonymized datasets should only be released under a license regime, rather than as fully open data. The higher the risk of de-anonymization, the more reason to tie a license to a dataset.

Access and re-use restrictions can also be combined. For instance, researchers could be required to visit the office of a public sector body to inspect a dataset: an access restriction. But at the same time, the

---

282. *Id.*

283. Solove makes a similar distinction between “access restrictions” and “use restrictions.” Solove, *supra* note 7, at 1169–70.

284. *Article 29 Opinion on Open Data and PSI Re-use*, *supra* note 203, at 25–26; *see also* NARAYANAN, HUEY & FELTEN, *supra* note 217, at 18. An issue that falls outside the scope of this paper is the legal basis for such licenses. In some countries, the public sector might have a type of intellectual property right on the dataset; in other countries the public sector body could invoke general contract law to impose a license on the dataset.

285. SID/SASD/SEDD Application Kit, HEALTHCARE COST & UTILIZATION PROJECT 24 (Sept. 16, 2015), [http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD\\_Final.pdf](http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf).

286. *Article 29 Opinion on Open Data and PSI Re-use*, *supra* note 203, at 25.

287. *Id.*

researchers could be required to not try to re-identify people in the dataset: a re-use restriction.

### C. DISCLOSURE AS OPEN DATA

The third access type is releasing data as fully open data: with no access or re-use restrictions. For instance, perhaps some personal data included in lobbying or company registers should be released as open data. Restricting access or re-use might make it too difficult to analyze the influence of lobbyists or to hold companies accountable.<sup>288</sup>

In conclusion, sometimes a balance can be struck between open data goals and privacy by disclosing data with access or re-use restrictions, rather than as fully open data. Hence, a public sector body must first assess whether a dataset should be disclosed at all. If it is decided that data should be disclosed, the next question is whether the data should be released with access or re-use restrictions, or as fully open data.

## VII. A CIRCUMSTANCE CATALOGUE TO INFORM DISCLOSURE DECISIONS

The above suggests that public sector bodies should decide on a case-by-case basis whether, and under which conditions, a dataset should be disclosed.<sup>289</sup> Narayanan et al. note that “[e]ach dataset has its own risk-benefit tradeoff, in which the expected damage done by leaked information must be weighed against the expected benefit from improved analysis.”<sup>290</sup> The researchers add that “[b]oth assessments are complicated by the unpredictable effects of combining the dataset with others, which may escalate both the losses and the gains.”<sup>291</sup>

There is not one clear-cut rule to decide whether datasets including or based on personal data should be disclosed. The lack of a hard-and-fast rule is not surprising. As discussed *supra* in Part III, the problem of

---

288. See, e.g., Jonathan Gray & Tim Davies, *Fighting Phantom Firms in the UK: From Opening Up Datasets to Reshaping Data Infrastructures?* (May 27, 2015) (Working paper presented at the Open Data Research Symposium, 3rd International Open Government Data Conference, Ottawa), <http://ssrn.com/abstract=2610937>; TRANSPARENCY INTERNATIONAL, *HOW OPEN DATA CAN HELP TACKLE CORRUPTION* (2015), <http://www.transparency.org.uk/publications/how-open-data-can-help-tackle-corruption-policy-paper>.

289. Many authors arrive at that conclusion. See, e.g., Katleen Janssen & Sara Hugelier, *Open Data: A New Battle in an Old War Between Access and Privacy?*, in *DIGITAL ENLIGHTENMENT YEARBOOK 2013*, at 190, 199 (Mireille Hildebrant et al. eds., 2013).

290. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 12.

291. *Id.*; see also *id.* at 13, 15.

balancing privacy and open data interests can be seen as a modern version of the problem of balancing privacy and public sector transparency.

The objectives behind open data policies and corresponding public interests involved merit closer scrutiny; this allows for differentiation that is necessary for balancing the interests involved.<sup>292</sup> The general FIPs guidance suggesting a balance between privacy and other interests is not detailed enough in the case of open data. We propose that a circumstance catalogue can help to decide whether and how to release data.<sup>293</sup> The circumstance catalogue lists circumstances, or factors, that should be considered when assessing whether, and under which conditions, a dataset should be released, as well as different options for how it should be released. We provide a list as a starting point for a debate—the list is not meant to be exhaustive or final. The circumstance catalogue can be extended, for instance, by taking inspiration from case law, freedom of information law, and guidelines regarding open data and privacy.

We mention some rules of thumb regarding re-identification risks and releasing data. One rule of thumb is that raw personal data should generally not be released as fully open data, unless there is a compelling public interest argument for choosing this route for disclosure over other available options.<sup>294</sup> We argue that pseudonymous data must generally be treated as a type of personal data, rather than as anonymous data. On the other hand, non-personal data can generally be released as open data. For purportedly anonymized data, it is more complicated. As stated previously, irreversible anonymization is difficult, and perhaps impossible to achieve.<sup>295</sup> Therefore, in some cases anonymized data should not be released as fully open data.

#### A. WEIGHT OF THE GOALS PURSUED

The goals pursued by disclosing data are relevant. The consideration is not only what the (theoretical) aim of the public body is. An assessment might also be made of the most likely uses of the data by other public bodies, the private sector, and citizens. True, this runs counter to the idea behind open data that serendipitous re-use is positive, and that it is

---

292. See Scassa, *supra* note 71, at 405.

293. For a similar approach, balancing interests in access to court records against other considerations, see Conley, Datta, Nissenbaum & Sharma, *supra* note 234, at 797–98.

294. Narayanan et al. reach a similar conclusion. NARAYANAN, HUEY & FELTEN, *supra* note 217, at 15 (“[I]t almost never will be the case that an unlimited release of a dataset to the entire public will be the optimal choice.”).

295. See *supra* Section V.E. See also Narayanan & Shmatikov, *supra* note 266, at 26.

impossible for the government to predict potential uses.<sup>296</sup> But it is naïve to assume that uses will all be benevolent.

What is the primary goal pursued with releasing data and how important is releasing this type of information, in this form, to achieving that goal? Could the objective be adequately addressed by disclosing information in a less privacy-sensitive form? Is it likely that the data will be used primarily by the press or similar public watchdogs, or are the data primarily interesting for commercial purposes?<sup>297</sup> The more relevant data are to key aspects of democratic participation, the stronger the case for release as open data. As Daniel Solove notes, when deciding whether to release personal data, political transparency has more weight than pure commercial interests of re-users:

Access should be granted for uses furthering traditional functions of transparency such as the watchdog function; access should be denied for commercial solicitation uses because such uses do not adequately serve the functions of transparency. Rather, such uses make public records a cheap marketing tool, resulting in the further spread of personal information, which is often resold among marketers.<sup>298</sup>

Furthermore, not all uses of public sector information are equal before the law. Additionally, the national legal system makes a difference. For instance, the strength of rights to access to information and the discretionary space for public authorities differ from country to country. For example, in the United States, the First Amendment influences decisions regarding data disclosure.<sup>299</sup> In Europe, access to information to foster political transparency also has backing in human rights treaties.<sup>300</sup> But in Europe, legal privacy and data protection rights have more relative

---

296. For example, the G8 Open Data Charter contains the pledge of governments to ensure “that the data are available to the widest range of users for the widest range of purposes.” G8 OPEN DATA CHARTER, *supra* note 1. Assessing the market for public sector information based products and services in the U.K., Deloitte concludes that “it is hard to foresee specifically where innovation might take place in the U.K. Often innovation takes place in areas which are hard to predict.” DELOITTE, MARKET ASSESSMENT OF PUBLIC SECTOR INFORMATION, *supra* note 38, at 41.

297. The U.S. Supreme Court noted that different purposes have different weights in the context of inspecting and copying judicial records. *Nixon v. Warner Commc’ns, Inc.*, 435 U.S. 589 (1978). In the FOIA context, the Supreme Court arrived at a similar conclusion. *U.S. Dep’t of Justice v. Reporters Comm. for Freedom of the Press*, 489 U.S. 749, 773 (1989).

298. Solove, *supra* note 7, at 1192.

299. *See id.* at 1200–06; *see also* *Sorrell v. IMS Health Inc.*, 131 S. Ct. 2653 (2011).

300. Mireille van Echoud & Katleen Janssen, *supra* note 117, at 483–88.

weight than in the United States.<sup>301</sup> An important factor in this respect is the role of people whose data are considered for release. Do the data concern somebody who holds a public function or a powerful position? What is the level of responsibility of the person? To what extent is the information needed in open, machine-readable form in order to facilitate democratic accountability? The higher the level of responsibility of the person, the more likely it is that transparency trumps privacy interests.

While access to information to foster democratic transparency has backing in constitutions and human rights documents, the legal backing of releasing information for business opportunities or for improving public sector efficiency is less evident.<sup>302</sup> If there is a good case for sharing data within the public sector because this contributes to efficient government, governments should regulate such sharing with specific laws that contain appropriate safeguards. Cost and efficiency savings in and of themselves may not outweigh the protection of individual privacy unless there are other overriding concerns about, for example, public accountability, corruption, or the exercise of democratic oversight.

Apart from the difference in national legal systems, the weight of the goal also depends on the national situation. For example, in a country where there are many problems with corruption by state officials, disclosing detailed wealth records of public functionaries makes more sense than in a country with virtually no corruption. And in some countries there may be more widespread acceptance of the public disclosure of salaries.<sup>303</sup>

## B. WEIGHT OF THE PRIVACY INTERESTS

Arguments against releasing data, or against releasing data without restrictions, include the following: there are considerable risks associated with releasing the data; the potential harm is serious, rather than a minor inconvenience; the privacy of many people (not a few) is at risk; and the privacy threat is immediate rather than remote. For instance, a theoretical privacy infringement has less weight than would a clear danger. A clear privacy danger might occur, for example, with a dataset containing names of people with HIV. People could be discriminated against if it becomes publicly known that they have HIV.

---

301. See generally Kranenborg, *supra* note 139.

302. The Charter of Fundamental Rights of the European Union does recognize the right to do business. E.U. Charter of Fundamental Rights, *supra* note 68, art. 16.

303. For instance, in Finland, the tax authorities disclose the income of people whose income exceeds certain thresholds. See Case C-73/07, *Tietosuojavaltuutettu v Satakunnan Markkinapörssi Oy*, 2008 E.C.R. I-9831.

The nature of the harm also matters: for example, if the data relate to people fulfilling public functions and concern professional conduct, a risk of reputational harm is unlikely to be of concern (unless there is doubt about the accuracy of the data). That would be the case with disclosing expenses claims. If disclosure leads to a security risk, e.g., disclosing an itinerary or detailed information about a politician's movements, the case is different.

Expectations of privacy can also be a factor. How were the data collected? If there was a promise or understanding of confidentiality, the case is different than if people have volunteered data after they were warned of future possible disclosures. Because of asymmetry in information relationships between public authorities and citizens, it cannot be readily assumed that data was truly volunteered.

In conclusion, a case-by-case analysis is required when deciding whether to release data, and whether the data can be disclosed as fully open data, or whether access or use should be restricted. We proposed a starting point for a circumstance catalogue that would help to assist in decisions about data disclosure.

## VIII. CONCLUSION

Open data are held to contribute to a wide variety of social and political goals—including strengthening transparency, public participation, and democratic accountability; promoting economic growth and innovation; and enabling greater public sector efficiency and cost savings. But releasing datasets as open data may threaten privacy, for instance if they contain personal or re-identifiable data. Potential privacy problems include chilling effects on people communicating with the public sector, a lack of individual control over personal information, and discriminatory practices enabled by the released data.

Can privacy and related interests be respected, while not unduly hampering open data benefits? The Fair Information Principles (FIPs), as expressed in the OECD Privacy Guidelines, provide a framework to balance privacy and other interests. From a FIPs perspective, the main problem with open data is that it can be used by anyone for any purpose. A complete lack of re-use restrictions would clash with the purpose specification principle of the FIPs. It follows from the purpose specification principle that personal data should only be collected for a purpose that is specified in advance, and that those data should not be used for incompatible purposes.

Compromises are possible to balance privacy and open data interests. We distinguish between four data categories with different risk levels: raw

personal data, pseudonymous data, anonymized data, and non-personal data. With raw personal, no attempt has been made to make identification harder. Pseudonymous data are data for which the individual's name is changed to another unique identifier. Anonymized data are ex-personal data; people cannot be re-identified in the dataset. Non-personal data, such as data about weather conditions or public transport times, never contain personal data.

Non-personal data can generally be released without restrictions as fully open data. As a rule of thumb, raw personal data should not be released as fully open data. Pseudonymous data must generally be treated as a type of personal data—not as anonymous data. Anonymized data is more complicated. Anonymized data can, in theory, be disclosed as open data, without re-use restrictions. However, irreversible anonymization is exceedingly difficult, and perhaps impossible. And even in aggregated and purportedly anonymized data, individuals can sometimes be re-identified. Therefore, some purportedly anonymized datasets should only be disclosed with access and re-use restrictions.

Sometimes, a compromise can be found by releasing anonymized data with access and re-use restrictions. Restricting openness can be done in various ways. For instance, the public sector body could attach a license to the data, requiring the re-user to only use certain data for a certain purpose (say medical research) and to promise not to re-identify the data. Other limitations on openness can also be envisaged. For instance, if a research interest is important, but the personal data are sensitive, researchers could be required to visit the lab where the data are held.

Hence, a case-by-case analysis is required when deciding whether to release data, and whether the data can be disclosed as fully open data, or whether access or use should be restricted. To assist in decisions about data disclosure, a circumstance catalogue may be of help: a list of circumstances to consider when deciding about releasing data. For instance: what is the goal pursued by releasing the data? Is there another way to pursue that goal? What are the risks involved with releasing the data? Are the privacy-related risks negligible or probable? If the risk materializes, what is the harm that results? Is the privacy of a few or of millions of people at stake?

In conclusion, in many instances public sector datasets that contain, or are based on, personal data should not be released as fully open data. When arguments for disclosure do outweigh privacy interests, open data should not be considered the only route. Other options might include disclosing information with access or re-use restrictions.

