TOWARDS A MODERN APPROACH TO PRIVACY-AWARE GOVERNMENT DATA RELEASES

Micah Altman, MIT; Alexandra Wood, David R. O'Brien, Salil Vadhan, Urs Gasser, Harvard University[1]

TABLE OF CONTENTS

1 Micah Altman and Alexandra Wood were 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 the funding acquisition in support of this research. The research and the writing of this report were supported by Microsoft Corporation in collaboration with the Berkeley Center for Law & Technology. In addition, this material is based upon work supported by the National Science Foundation under Grant No. 1237235 as well as the Ford Foundation and the John D. and Catherine T. MacArthur Foundation. We wish to thank the members of the Privacy Tools for Sharing Research Data project for helpful comments.

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. Furthermore, releases of government data promote growth in the private sector, guiding investment and other commercial decisions, supporting innovation in the technology sectors, and promoting economic development and competition generally.[2] 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.[3] 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.[4] 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 queries, is a widely publicized example of the unexpected uses of data.[5] These are, of course, just a few examples of the many benefits of open data.[6]

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.[7] 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.[8] 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.[9]

---

2 *See generally* REGINA POWERS & DAVID BEEDE, U.S. DEP'T OF COMMERCE, FOSTERING INNOVATION, CREATING JOBS, DRIVING BETTER DECISIONS: THE VALUE OF GOVERNMENT DATA (2014) (discussing the many benefits of government releases of data).

3 *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).

4 *See* sources cited *supra* note 3.

5 *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://doi.org/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://doi.org/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.

6 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).

7 *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-.

8 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).

9 *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.

However, a major challenge for any public release of data about individuals is providing meaningful protection of privacy interests.[10] While governments are generally required to consider the legal and ethical implications of publicly releasing information about individuals, the disclosure and reuse of privacy-sensitive data are greatly hindered by the lack of an effective legal and regulatory framework for privacy.[11] Privacy laws and policies can be circumstantial, open to interpretation, and ill-suited to apply at scale.[12] Most states lack "omnibus data protection laws" and have "scattered laws [that] provide only limited protections for personal information in the public sector."[13] Instead, 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.[14] Executive agencies, for example, frequently release government information under the Freedom of Information Act (FOIA),[15] 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, 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.[16] These agencies use statistical disclosure limitation techniques to aggregate information from many individuals, suppress sensitive individual-level details, or perturb individual data points in ways intended to mitigate privacy concerns while supporting accurate analyses.[17]

As numerous commentators have shown, naïve treatment of information privacy and security has become a major stumbling block to efficient access to and use of data.[18] Assessment of privacy risk

---

10 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" to refer to the right of data subjects to control the manner and extent of sharing of their information and "confidentiality" to refer to the duty of data holders to prevent unauthorized disclosure after collection. *See, e.g.*, Stephen E. Fienberg, *Confidentiality and Disclosure Limitation*, 1 ENCYCLOPEDIA OF SOCIAL MEASUREMENT 463 (2005). In contrast, the literature on cryptography often uses "privacy" to refer to controls over disclosure or to the absence of a privacy breach, *see, e.g.*, Cynthia Dwork, *Differential Privacy*, ENCYCLOPEDIA OF CRYPTOGRAPHY AND SECURITY 338 (2011), and 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 (2006).

11 *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").

12 *Id.*; Paul Schwartz, *Data Processing and Government Administration: The Failure of the American Legal Response to the Computer*, 43 HASTINGS L.J. 1321 (1992).

13 Schwartz, *supra* note 8, at 605.

14 *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).

15 Freedom of Information Act, 5 U.S.C. § 552 (2013).

16 Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, 116 Stat. 2899, 2962 (2002) (codified at 44 U.S.C. § 3501 note (2013)).

17 FED. COMM. ON STATISTICAL METHODOLOGY, STATISTICAL POLICY WORKING PAPER 22 (SECOND VERSION), REPORT ON STATISTICAL DISCLOSURE LIMITATION METHODOLOGY (Dec. 2005), https://fcsm.sites.usa.gov/files/2014/04/spwp22.pdf.

18 *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

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.[19] Yet privacy risks and harms are difficult to predict as data are accumulated, combined, and used in a wide variety of contexts,[20] and data release programs often fail to address 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 common practices such as removing or generalizing sensitive fields.[21] In addition, these techniques significantly reduce the utility of data.[22] On the whole, robust 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.[23] These issues are 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.[24]

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.[25] 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.[26] Yet current laws and policies do not provide much guidance to agencies regarding the implementation of stronger privacy protections in their public releases of data.[27] Taken

---

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).

19 *See* discussion *infra* Part III.

20 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/htm.

21 *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, 3:26 AM), http://www.haaretz.com/news/national/israel-s-anonymous-statistics-surveys-aren-t-so-anonymous-1.492256.

22 *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.

23 *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).

24 *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 23; 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).

25 *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://www.dataprivacylab.org/projects/irb/Vadhan.pdf.

26 *See id.*; *see also, e.g.*, 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 CONFERENCE ON DATA ENGINEERING 277 (2008), http://www.cse.psu.edu/~dkifer/papers/PrivacyOnTheMap.pdf.

27 *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 (Apr. 2001), http://www.gao.gov/new.items/d01126sp.pdf.

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 advances in data privacy from disciplines including computer science,[28] statistics,[29] and law,[30] 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 mechanism.

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[31] 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, among other things, 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.

Government agencies attempt to protect the privacy of individuals whose information may be present in these date releases. For example, an agency might redact certain identifiers such as first and last names or might withhold the release of a record entirely. In some cases an agency is bound by regulations requiring strong confidentiality protections for collecting and releasing information about individual respondents,[32] while in other cases an agency may not be required to prevent the release of personal information at all. 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.

28 *See, e.g.,* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMUNICATIONS OF THE ACM 86 (2011); Erica Klarreich, *Privacy by the Numbers: A New Approach to Safeguarding Data*, SCIENTIFIC AMERICAN QUANTA MAGAZINE, Dec. 31, 2012; Ori Heffetz & Katrina Ligett, *Privacy and Data-Based Research*, 28 J. ECON. PERSPECTIVES 75 (2014).
29 *See, e.g.*, Machanavajjhala et al., *supra* note 26.
30 *See, e.g.*, Wu, *supra* note 24.
31 *See* 5 U.S.C. § 552 (2013).
32 *See, e.g.*, Confidentiality Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, 116 Stat. 2899, 2962 (2002) (codified at 44 U.S.C. § 3501 note (2013)).

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 U.S. 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:[33] responses to freedom of information and Privacy Act[34] requests,[35] traditional public and vital records,[36] official government statistics,[37] and e-government and open government initiatives.[38] 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.

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.[39] In combination, these laws are intended to strike a balance between the public's right to know what information is held by the government and the government's interest in safeguarding sensitive information the release of which could harm protected individual, commercial, or governmental interests.[40]

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.[41] The FOIA process is used very frequently, with requests across all federal agencies totaling 714,231 in 2014.[42] FOIA empowers any person, including a non-citizen, to obtain copies of records held by federal executive agencies by following a simple request procedure.[43] 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

33 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.
34 Privacy Act of 1974, 5 U.S.C. § 552a (2013).
35 *See infra* Section II.A.1.
36 *See infra* Section II.A.2.
37 *See infra* Section II.A.3.
38 *See infra* Section II.A.4.
39 *See, e.g.*, Government in the Sunshine Act, 5 U.S.C. § 552b (2013) (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").
40 *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).
41 *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).
42 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).
43 *See, e.g.*, U.S. Dep't of Justice v. Reports Comm. for Freedom of the Press, 489 U.S. 749 (1989).

commercial reasons.[44] 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 authorized by statute.[45] By default, all responsive records must be disclosed upon request unless an applicable exemption, such as privacy,[46] 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. For example, a New York state law 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.[47] Some state freedom of information laws also expressly allow victims of crimes to shield their personal information from release.[48] Otherwise, 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. Furthermore, once released, the information can be used for any purpose and freely redisseminated, and no efforts are made to monitor access to the data or mitigate threats at the post-access stage. 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.[49]

A companion law, the Privacy Act of 1974,[50] 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 that is maintained in a system of records, except as authorized by the data subject.[51] It authorizes FOIA-mandated disclosures,[52] 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.[53] The Privacy Act also enables a data subject to access,

---

44 *See* Cate et al., *supra* note 41, 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).

45 *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).

46 5 U.S.C. § 552(b)(6), (b)(7) (2013).

47 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.

48 *See, e.g.*, CAL. GOV'T CODE § 6254(f)(2) (West 2015).

49 5 U.S.C. § 552(a)(4).

50 5 U.S.C. § 552a.

51 5 U.S.C. § 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," 5 U.S.C. § 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," 5 U.S.C. § 552a(a)(4).

52 5 U.S.C. § 552a(b)(2).

53 *See* Savada v. U.S. Dep't of Defense, 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, 819 F.2d 1181, 1184 (D.C. Cir.

review, and correct her information in government databases, unless an exemption applies.[54] An individual may submit a written Privacy Act request to access records about herself.[55] 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.[56]

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 prompted 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.[57] The Centers for Medicare and Medicaid Services disclosed payments made by pharmaceutical companies to individual doctors and the brand names and quantities of medications the doctors prescribed, and the data were published by a watchdog group in online searchable databases, along with visualizations and investigative commentary on prescribing patterns and signs of fraud.[58] During a public debate about gun control legislation, a newspaper used freedom of information requests to obtain county agencies' data, from which it created a widely publicized interactive map showing the names and addresses of handgun permit holders.[59] 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.[60]

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 it can disseminate data produced by the research in response to FOIA requests.[61]

FOIA exempts the following records from mandatory release: classified records; internal personnel records and agency memos; confidential trade secret or financial information; medical or other similar files, broadly interpreted,[62] that would constitute an unwarranted invasion of privacy; and

---

1987) ("If a FOIA exemption covers the documents, but a Privacy Act exemption does not, the documents must be released under the Privacy Act.")).

54 5 U.S.C. § 552a(d)(1)–(2).

55 5 U.S.C. § 552a(d)(1).

56 5 U.S.C. § 552a(g)(1).

57 *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.

58 *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).

59 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.

60 *See LoHud Removes Controversial Gun Owners Map*, NBC 4 NEW YORK, Jan. 18, 2013, http://www.nbcnewyork.com/news/local/Journal-News-Removes-Pistol-Permit-Database-Gun-Owners-Rockland-Westchester-187525461.html (includes a video showing the features of the original interactive map).

61 Shelby Amendment, Pub. L. No. 105-277, div. A, tit. III, 112 Stat. 2681, 2681-495 (1999).

62 U.S. Dep't of State v. Washington Post Co., 456 U.S. 595, 599–603 (1982).

law enforcement records; among several other categories.[63] Agencies are permitted but not required to withhold or redact records that fall within one of the exemptions,[64] and they are generally encouraged to release exempted information, when possible, "as a matter of good public policy."[65] 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."[66]

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."[67] Federal courts have applied different tests for determining whether a particular piece of information falls within this definition,[68] 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.[69] However, it can still 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.[70] An agency may also disclose information for any "routine use" that is "compatible" with its purpose for collecting the information.[71] Commentators have argued that this provision effectively enables disclosure with very little restriction.[72]

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"[73] and where the threats to

63 5 U.S.C. § 552b(c) (2013).

64 *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").

65 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.

66 CAL. GOV'T CODE § 6255 (West 2015).

67 Responsibilities for the Maintenance of Records About Individuals by Federal Agencies, 40 Fed. Reg. 28,948, 28,951-52 (July 9, 1975).

68 *Compare, e.g.*, 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, e.g.*, 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").

69 "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.

70 5 U.S.C. § 552a(b)(4), (b)(7), (b)(9), (b)(12) (2013).

71 5 U.S.C. § 552a(b)(3).

72 *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).

73 *See* Dep't of the Air Force v. Rose, 425 U.S. 352, 382 (1976).

privacy are "more palpable than mere possibilities."[74] Yet it has also held that records need not contain "highly personal" information or "intimate details" to be considered to be privacy-sensitive.[75] 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."[76] 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."[77] 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.[78]

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.[79] In one 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.[80] 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."[81] 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.[82] If a request is drawn narrowly such that responding to it 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.[83]

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.[84] Rather, case-by-case determinations regarding the information to withhold or release in

74 *Id.* at 380 n.19.
75 U.S. Dep't of State v. Washington Post Co., 456 U.S. 595, 600–01 (1982).
76 *See* U.S. Dep't of Justice v. Reporters Comm. for Freedom of the Press, 489 U.S. 749, 763–74, 774, 780 (1989).
77 U.S. Dep't of State v. Ray, 502 U.S. 164, 176–77 (1991).
78 U.S. Dep't of Def. v. Fed. Labor Relations Auth., 510 U.S. 487, 500–01 (1994) (emphasis omitted from first quotation).
79 *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).
80 Jorge Fitz-Gibbon, *Putnam Must Release Gun Records, Judge Says*, THE JOURNAL NEWS (White Plains, N.Y.) (Mar. 5, 2014, 11:16 PM), http://www.lohud.com/story/news/2014/03/05/journal-news-putnam-gun-map-lawsuit/6097983.
81 Arieff v. U.S. Dep't of Navy, 712 F.2d 1462, 1467 (D.C. Cir. 1983) (emphasis in original).
82 *See, e.g.*, Multi Ag Media LLC v. U.S. Dep't of Agric., 515 F.3d 1224, 1230 (D.C. Cir. 2008).
83 *See* Dep't of the Air Force v. Rose, 425 U.S. 352, 381 (1976); *see also, e.g.*, 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).
84 *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).

response to a FOIA request often vary according to the "position, background, and training" of the official making the decision.[85]

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.[86] 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.[87] 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.[88] 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.[89]

The Privacy Act's redress mechanisms are widely considered weak. To enforce her rights, 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.[90] Even then, the Act limits potential remedies to injunctions requiring an agency to correct the 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.[91] 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."[92] 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."[93]

Freedom of information laws are a burdensome mechanism for releasing information. Freedom of information decisions are discretionary, the management of requests and compliance is

85 Lotte E. Feinberg, *Managing the Freedom of Information Act and Federal Information Policy*, 46 PUB. ADMIN. REV. 615, 617 (1986).
86 *See* 5 U.S.C. § 552(b) (2013).
87 *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).
88 *Compare, e.g.*, 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*, *e.g.*, 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).
89 *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.
90 *See, e.g.*, BeVier, *supra* note 84, at 479–82.
91 5 U.S.C. § 552a(g)(2)–(4).
92 Schwartz, *supra* note 8, at 596.
93 *See* Andrews v. Veterans Admin., 838 F.2d 418, 421, 425 (10th Cir. 1988).

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.[94] 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.[95] 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.[96] In 2012, the government launched FOIAonline,[97] 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.[98]

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.[99] 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.[100] 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.[101] 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.[102]

---

94 In 2014, the total backlog of FOIA requests across the federal government was 159,741. U.S. Dept. of Justice, *supra* note 42.

95 Chris Whong, *FOILing NYC's Taxi Trip Data*, BLOG (Mar. 18, 2014), 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).

96 *See, e.g.*, *FOIA Library*, U.S. CENSUS BUREAU, http://www.census.gov/about/policies/foia/foia_library.html (last modified Oct. 10, 2014).

97 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.

98 *See infra* Section II.A.4.

99 *See* Daniel J. Solove, *Access and Aggregation: Public Records, Privacy, and the Constitution*, 86 MINN. L. REV. 1137, 1173–76 (2002).

100 *See id.*

101 *See* ProtectMarriage.com v. Bowen, 752 F.3d 827, 835 (9th Cir. 2014).

102 *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.

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.[103] Historically, there were practical barriers limiting access to vital and public records, such as the necessity of visiting a local office in person during regular business hours to physically search and inspect available records.[104] Locating records of interest through this process could involve trips to multiple offices and significant expenditures of time and money. In addition, 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 been increasingly 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.[105] 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 and have registered for remote access and signed a subscription agreement.[106] 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. LexisNexis, for example, provides a database for mining over 36 billion public records collected from state agencies.[107]

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

103 *See generally* David R. O'Brien et al., *Integrating Approaches to Privacy Across the Research Lifecycle: When Is Information Purely Public?* (Working Paper, Mar. 27, 2015), http://www.ssrn.com/abstract=2586158 (discussing the gap between expectations of privacy and the increasing public availability of personal information).

104 *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.").

105 *See, e.g.*, *Remote Access Site*, CITY OF CHESAPEAKE CLERK OF CIRCUIT COURT, http://www.chesapeakeccland.org (last visited Apr. 28, 2015) ("providing access to land and other related records maintained by this office").

106 *See Access to Case Information*, RHODE ISLAND JUDICIARY, https://www.courts.ri.gov/Pages/access-caseinfo.aspx (last visited July 13, 2015).

107 *See* Brochure, LexisNexis, Ten Compelling Reasons to Rely on LexisNexis Public Records as You Research People, Businesses, and Locations (2012), http://www.lexisnexis.com/pdf/Ten Reasons_Corp_Gov_FINAL.pdf.

contain information, such as size and assessment value, that can reflect the owner's financial situation. Arrest records[108] 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.[109] 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).[110]

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."[111] 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."[112] State and local public records laws arguably provide weak protection for individual privacy,[113] 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[114] and requiring individuals seeking public records to pledge not to use the information for solicitation or marketing.[115] Federal law also restricts the disclosure of state public records in a few narrow categories. The Driver's Privacy Protection Act,[116]

---

108 *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).

109 *See, e.g.*, Maine State Police, Maine Sex Offender Registry, 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).

110 *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.").

111 Nixon v. Warner Commc'ns, Inc., 435 U.S. 589, 598 (1978) (citations omitted).

112 *Id.* at 599.

113 *See, e.g.*, Solove, *supra* note 99, at 1154–72.

114 *See, e.g.*, VA. CODE ANN. § 46.2-208 (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.").

115 *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").

116 Driver's Privacy Protection Act of 1994, 18 U.S.C. § 2721 (2013).

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 such as the Family Educational Rights and Privacy Act (FERPA)[117] and the Health Insurance Portability and Accountability Act (HIPAA)[118] prohibit the release of certain education and health care records, respectively. Outside of these narrow restrictions, public records are generally made freely 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 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."[119] 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.[120] In addition, it required the records to be released almost in full, subject only to redaction of the names of the individuals receiving services.[121]

Many examples demonstrate the difficulty of making release decisions and adequately safeguarding personal information when bound by state public records laws. For example, in 2003, the county clerk for a Virginia court digitized many of the court's public records to make them available online.[122] 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.[123] 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.[124] 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.[125]

---

117 Family Educational Rights and Privacy Act of 1974, 20 U.S.C. § 1232g (2013).
118 HIPAA Privacy Rule, 45 C.F.R. pt. 160 and subpts. A & E of pt. 164 (2014).
119 *See* FED. R. CIV. P. 26(c).
120 State Dep't of Pub. Health v. Superior Court, 342 P.3d 1217 (Cal. 2015).
121 *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.")).
122 *See* Dan Telvock, *Board Passes Resolution to Delay Remote Access of Public Court Records that Contain Personal Data*, LEESBURG TODAY, July 21, 2003.
123 *Id.*
124 *Id.*
125 *Id.*

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.[126] 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.[127] The statistics it produces are used to draw political districts, apportion seats in the U.S. House of Representatives, distribute federal funds across the country, and guide the decisions of governments and businesses, among many other uses.[128]

Statistical agencies employ strict confidentiality protections, backed by federal laws such as the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA),[129] to maintain public trust, ensure data integrity, and promote the sustainability of statistical programs.[130] A key privacy threat is the identification of an individual in published data, which typically is a violation of law and threatens public confidence in statistical agencies' collection and analysis of personal information.[131] 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.[132] Another threat is inappropriate integration of different types of data across multiple government organizations, which is legally constrained and bounded in part by the public's expectations about how the government uses their personal information and general concerns about government surveillance.[133] Commercial firms are also concerned about official statistics leaking their competitive information.[134]

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

126 *See, e.g.*, *Databases, Tables & Calculators by Subject*, U.S. BUREAU OF LABOR STATISTICS, http://www.bls.gov/data (last visited May 26, 2015).
127 *See* Lawrence H. Cox & Laura V. Zayatz, *An Agenda for Research in Statistical Disclosure Limitation*, 11 J. OFFICIAL STATISTICS 205 (1995).
128 *See* U.S. CENSUS BUREAU, MEASURING AMERICA: THE DECENNIAL CENSUSES FROM 1790 TO 2000 (Sept. 2002), https://www.census.gov/history/pdf/measuringamerica.pdf.
129 Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, 116 Stat. 2899, 2962 (2002) (codified at 44 U.S.C. § 3501 note (2013)).
130 *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) (Oct. 2006), https://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/proposed_cispea_guidance.pdf.
131 *See id.*
132 *See, e.g.*, Sweeney, *supra* note 21 (describing a record linkage attack on de-identified health data using public sources).
133 *See* Stephen E. Fienberg, *Toward a Reconceptualization of Confidentiality Protection in the Context of Linkages with Administrative Records*, 3 J. PRIVACY & CONFIDENTIALITY 65 (2011).
134 *See* Kinney et al., *supra* note 26.

characteristics such as age, gender, race, and ethnicity of the total population of the United States.[135] 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.[136] 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.[137]

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.[138] In some cases, agencies also make available more complex derived tables and, less frequently, geographically aggregated data or sanitized microdata. [139]

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.[140] 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 use this information to the detriment of the data provider."[141] Specifically, CIPSEA protects data collected for statistical purposes by a pledge of confidentiality to the respondent.[142] As required by CIPSEA, statistical agencies review data prior to release to ensure they do not contain information in identifiable form.[143] 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

135 *See* U.S. CENSUS BUREAU: 2010 CENSUS SUMMARY FILE 1: TECHNICAL DOCUMENTATION (Sept. 2012), http://www.census.gov/prod/cen2010/doc/sf1.pdf.

136 *See, e.g.*, *Databases, Tables & Calculators by Subject*, *supra* note 126.

137 *See, e.g.*, THOMAS D. SNYDER & SALLY A. DILLOW, U.S. DEPT. OF EDUC., NCES 2015-011, DIGEST OF EDUCATION STATISTICS 2013 (May 2015), http://nces.ed.gov/programs/digest/d13.

138 *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).

139 *See generally id.*

140 Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, 116 Stat. 2899, 2962 (2002) (codified at 44 U.S.C. § 3501 note (2013)).

141 *See* Margo Anderson & William Seltzer, *Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues*, 1 J. PRIVACY & CONFIDENTIALITY 7, 8 (2009).

142 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), 116 Stat. at 2963.

143 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), 116 Stat. at 2962.

confidentiality.[144] 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[145] 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.[146]

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."[147] Other laws may apply to the statistical activities of particular agencies such as the Internal Revenue Service[148] and the Social Security Administration.[149] For example, Title 13 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,"[150] and prohibits the use of statistical data for any purposes other than the statistical purposes for which it was supplied.[151] 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,"[152] that identifiable information are 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.[153] 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.[154] Agencies make public use data sets available

144 *See, e.g.*, CENSUS BUREAU, DISCLOSURE REVIEW BOARD (2001), https://www.census.gov/srd/sdc/wendy.drb.faq.pdf.

145 Disclosure risk refer 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.

146 *See id.*

147 Privacy Act of 1974, 5 U.S.C. § 552a(b)(5) (2013).

148 26 U.S.C. § 6108(c) (2013) ("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.").

149 42 U.S.C. § 1306(e)(3) (2013) ("[S]uch reports shall not identify individual patients, individual health care practitioners, or other individuals.").

150 13 U.S.C. § 9(a)(2) (2013).

151 13 U.S.C. § 9(a)(1).

152 72 Fed. Reg. 33362, 33371 (June 15, 2007).

153 *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 17 (providing an overview of statistical disclosure limitation techniques such as perturbation, aggregation, and suppression).

154 *See* FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 17.

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.

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.[155] Some agencies conduct background investigations on prospective researchers and hold them to the same confidentiality standards, backed by criminal penalties, as agency employees.[156] Researchers' use of restricted data is limited to the purposes they specified, and access is restricted to that necessary for the proposed analysis. [157] 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.[158] Agencies also employ technical controls on access and use via research data centers or enclaves[159] or, less frequently, remote analysis servers, which allow access to dynamically derived tables and maps.[160] 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.[161] Statistical agencies often evaluate the effectiveness of their disclosure limitation techniques by performing privacy impact assessments and staging reidentification attacks using available auxiliary data sets. For data users whose needs are not met by the public use data files, an agency may have a program to generate custom tabulations and review them by a disclosure review board before releasing them.[162]

Emerging challenges in this area include the rising speed of data collection and processing (sometimes referred to as data velocity),[163] heightened data integration,[164] and increasing analytical sophistication.[165] 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.[166] Agencies are pressured to release data faster, more cheaply, and in a way that allows a greater range of analysis,

---

155 *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 (describing the process of applying for access to research data through the Census Bureau Research Data Center).

156 *See, e.g.*, *id.* at 12.

157 *See, e.g.*, *id.* at 12-13.

158 *See, e.g.*, *id.*

159 *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).

160 *See, e.g.,* Michael Freiman et al., *The Microdata Analysis System at the U.S. Census Bureau*, JOINT STATISTICAL MEETINGS SECTION ON SURVEY RESEARCH METHODS (2011) (discussing a Census Bureau remote analysis server currently in development and providing an overview of similar systems that have been proposed or implemented).

161 *See, e.g.*, *OnTheMap*, http://onthemap.ces.census.gov (last visited May 28, 2015).

162 *See, e.g.*, *Special Tabulations Program*, https://www.census.gov/population/www/cen2000/sptabs/main.html (last visited May 28, 2015).

163 *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (May 2014), https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

164 *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 133.

165 *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.

166 *See, e.g.*, Carl Bialik, *Census Bureau Obscured Personal Data—Too Well, Some Say*, WALL ST, J. (Feb. 6, 2010, 12:01 AM), http://www.wsj.com/articles/SB10001424052748704533204575047241321811712.

including visualizations and data mining, and provides estimates for finer time scales and geographic areas.[167]

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."[168] 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."[169] 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 data releases.

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.[170] 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.[171] 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.[172] 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.[173] It also mandated that agencies identify additional high-value information and prepare a timeline for publishing this information online in open formats.[174] 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.[175] Finally, in 2013, President Obama signed an executive order directing the implementation of an Open Data Policy across the federal government based on the ideas 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,

167 *See* William E. Winkler, *Producing Public-Use Microdata that Are Analytically Valid and Confidential* (Census Bureau, Research Report No. RR98/02, 1998), https://www.census.gov/srd/papers/pdf/rr9802.pdf.
168 ORSZAG, *supra* note 9; *see, e.g.*, NYC OPEN DATA, https://nycopendata.socrata.com (last visited June 29, 2015).
169 ORSZAG, *supra* note 9.
170 OFFICE OF MGMT. & BUDGET, EXEC. OFFICE OF THE PRESIDENT, E-GOVERNMENT STRATEGY (Feb. 27, 2002), https://www.whitehouse.gov/sites/default/files/omb/inforeg/egovstrategy.pdf.
171 *Id.* at 1–2.
172 ORSZAG, *supra* note 9.
173 *Id.* at 7–8.
174 *Id.*
175 WHITE HOUSE, 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.

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."[176] The executive order also requires agencies to "safeguard individual privacy, confidentiality, and national security" when implementing the policy.[177]

a) Types of information released

The public release of data plays an essential role in these initiatives, and the 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.[178] 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.[179] In addition, making 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.[180] 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.[181]

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.[182] As of May 2015, 83 agencies and sub-agencies have published over 130,000 datasets on Data.gov,[183] though many of the datasets were published by just a handful of agencies or are duplicates of datasets previously posted elsewhere

176 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-.

177 *Id.*

178 *See, e.g.*, Daniel Tumminelli O'Brien, Robert J. Sampson, & Christopher Winship, *Ecometrics in the Age of Big Data: Measuring and Assessing "Broken Windows" Using Administrative Records*, Boston Area Research Initiative Working Paper No. 3 (2013).

179 See Steven E. Koonin, The Center for Urban Science and Progress,: The Promise of Urban Informatics (2013), http://cusp.nyu.edu/wp-content/uploads/2013/07/CUSP-overview-May-30-2013.pdf.

180 STEPHEN GOLDSMITH & SUSAN CRAWFORD, THE RESPONSIVE CITY: ENGAGING COMMUNITIES THROUGH DATA-SMART GOVERNANCE 78–79 (2014).

181 *See* TriMet, *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.

182 DATA.GOV, http://www.data.gov (last visited May 6, 2015).

183 *Id.*

online.[184] 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.[185] 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.[186]

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 portals.[187] 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.[188] 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 individual New York City apartment buildings.[189]

At the same time, the release of these data has privacy implications. For instance, 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.[190] 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. 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.

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. When collecting, storing, and sharing data about individuals, federal executive agencies must follow certain data security practices

184 Alon Peled, *When Transparency and Collaboration Collide: The USA Open Data Program*, 62 J. AM. SOC. FOR INFO. SCI. & TECH. 2085, 2088 (2011).

185 Todd Park & Steven VanRoekel, *Introducing: Project Open Data*, WHITE HOUSE OFF. SCI. & TECH. POL'Y BLOG (May 16, 2013, 9:46 AM), https://www.whitehouse.gov/blog/2013/05/16/introducing-project-open-data.

186 *See, e.g.*, City of Philadelphia, PHL API, 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).

187 *See* DATA.GOV, OPEN GOVERNMENT, https://www.data.gov/open-gov (last visited May 26, 2015).

188 GOLDSMITH & CRAWFORD, *supra* note 180, at 78.

189 *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.

190 *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.

prescribed by the National Institute of Standards and Technology,[191] disclosure limitation practices outlined in the 2005 Federal Committee on Statistical Methodology report,[192] and information privacy provisions in laws such as the Privacy Act of 1974,[193] the E-Government Act of 2002 (including CIPSEA),[194] and the Federal Information Security Management Act.[195] 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.)"[196] 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."[197] 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 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."[198] 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."[199]

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.[200] 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."[201] The third level, "non-public,"

191 *See, e.g.*, U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, STANDARDS FOR SECURITY CATEGORIZATION OF FEDERAL INFORMATION AND INFORMATION SYSTEMS, Federal Information Processing Standard (FIPS) Publication 199 (Feb. 2004) hereinafter U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, STANDARDS], http://csrc.nist.gov/publications/fips/fips199/FIPS-PUB-199-final.pdf; U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS, Special Publication 800-53, revision 4 (Apr. 30, 2013) [hereinafter U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, CONTROLS (draft)], http://csrc.nist.gov/publications/drafts/800-53-rev4/sp800-53-rev4-ipd.pdf.

192 FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY, *supra* note 17.

193 5 U.S.C. § 552a (2013).

194 E-Government Act of 2002, Pub. L. No. 107-347, 116 Stat. 2899.

195 Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–3549 (2013).

196 ORSZAG, *supra* note 9.

197 OFFICE OF MGMT. & BUDGET, OPEN DATA POLICY—MANAGING INFORMATION AS AN ASSET, MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES (May 9, 2013), https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

198 *Id.* at 9–10.

199 *Id.* at 10.

200 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).

201 *Id.*

is used for data that cannot be made available to the public and may only be shared within the federal government.[202]

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 uncertainty regarding regulatory requirements and that government lawyers frequently overinterpret legal standards.[203] For example, government employees may express a concern that privacy laws protect data held by their departments, but they lack guidance for screening specific datasets for release. Due to the existence of a specific privacy law, they 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.[204] 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.

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.[205] 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.[206] 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.[207] The checklist asks whether the dataset has previously undergone a formal disclosure committee review,[208] whether the data were collected from respondents under a promise of confidentiality, and whether a FOIA exemption applies to the information.[209] If the dataset contains microdata (individual-level rather than aggregate information), the checklist asks whether the microdata include direct identifiers ("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

202 *Id.*
203 GOLDSMITH & CRAWFORD, *supra* note 180, at 164–65.
204 GOLDSMITH & CRAWFORD, *supra* note 180, at 164–65.
205 DATA.GOV, NATIONAL/HOMELAND SECURITY AND PRIVACY/CONFIDENTIALITY CHECKLIST AND GUIDANCE, http://www.data.gov/sites/default/files/attachments/Privacy and Security Checklist.pdf (last viewed June 29, 2015).
206 *Id.* at 1–2.
207 *Id.* at 2.
208 *Id.* at 6.
209 *Id.* at 7.

business").[210] 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.[211] 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).[212] Indirect identifiers typically include other dates, locations and geographic information, and demographic characteristics such as gender or age.[213]

Agencies have established disclosure review practices for releasing 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,[214] the National Institute of Standards and Technology,[215] and Controlled Unclassified Information program[216] documents.[217] For instance, the E-Government Act of 2002 requires federal executive agencies to perform privacy impact assessments for their electronic information systems and any identifiable information about individuals they contain.[218] The Act direct 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.[219] 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.[220] 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.[221]

B. SHORTCOMINGS IN CURRENT PRACTICES

The foregoing discussion of many of the common approaches to releasing government data to the public 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 three commonly occurring shortcomings in privacy analysis and

210 *Id.* at 8.
211 *Id.* at 10.
212 *Id.* at 12.
213 *Id.* at 13–14.
214 OFFICE OF MGMT. & BUDGET, *supra* note 197.
215 U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, STANDARDS, *supra* note 191.
216 Exec. Order No. 13,556, 3 C.F.R. 267 (2011).
217 PROJECT OPEN DATA, *supra* note 200.
218 OFFICE OF MGMT. & BUDGET, OMB GUIDANCE FOR IMPLEMENTING THE PRIVACY PROVISIONS OF THE E-GOVERNMENT ACT OF 2002, Memorandum M-03-22 (Sept. 26, 2003), https://www.whitehouse.gov/omb/memoranda_m03-22.
219 *Id.*
220 *Id.*
221 *Id.*

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 that government data releases address, the approach that most government actors take 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.[222] Most state agencies similarly protect privacy by withholding categories of records, such as juvenile court records, or identifiable information in records, such as the names of sexual assault victims in police records, that are deemed to be sensitive.[223] 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 an information privacy law.[224]

This focus on a small set of controls appears suboptimal. It is now a well-established principle in the privacy science 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.[225] 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.[226] 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.[227] 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 using only 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, the National Institute of Standards and Technology 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."[228] General guidance directing agencies to protect the privacy of

222 *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.

223 *See* discussion *supra* Section II.A.2.

224 *See* discussion *infra* Section IV.B.

225 *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 INTERNATIONAL INT'L JOURNAL J. OF UNCERTAINTY FUZZINESS AND & KNOWLEDGE-BASED SYSTEMS 557 (2002).

226 *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).

227 *See id.*

228 NIST, PRIVACY RISK MANAGEMENT FOR FEDERAL INFORMATION SYSTEMS 1, Internal Report 8062 (Draft) (May 2005), http://csrc.nist.gov/publications/drafts/nistir-8062/nistir_8062_draft.pdf.

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 paper revealed only a handful of well-recognized or widely adopted sources on identifying and mitigating privacy risks.[229] In addition, on the whole this formal guidance is general, abstract, infrequently updated, and self-directed.[230] 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.[231] In contrast, formal guidance for analyzing and mitigating related information security risks, such as that described in FISMA,[232] is voluminous, proscriptive, specific, actionable, frequently updated, and integrative into legal systems of audit and certification.[233] The comparative paucity of privacy documentation often leads to inconsistent identification of privacy risks and ineffective application of privacy safeguards.[234]

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 to either 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.[235] Examples in the latter category include state freedom of information laws that expressly require redaction of identifying information about sexual assault victims.[236] 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

229 *See, e.g.*, FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY, *supra* note 17; U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS, NIST Special Publication 800-53 (Apr. 2013) [hereinafter U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, CONTROLS (final)], http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf; NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, GUIDE TO PROTECTING THE CONFIDENTIALITY OF PERSONALLY IDENTIFIABLE INFORMATION (PII), NIST Special Publication 800-122 (April 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 (July 1973), http://www.justice.gov/opcl/docs/rec-com-rights.pdf.

230 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, *see* FED. COMM. ON STATISTICAL METHODOLOGY, *supra* note 17, was last revised in 2005. 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.*

231 *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).

232 Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–49 (2013).

233 *See* discussion *infra* Section III.B.

234 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.

235 5 U.S.C. § 552(b)(6) (2013).

236 *See, e.g.*, Colo. Rev. Stat. § 24-72-304(4) (2011).

identifying particular assigned to the individual,"[237] and statistical agencies are likewise prohibited from disclosing information about individuals in identifiable form.[238] 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, it 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.[239] 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.

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 largely 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. Proposals for intervention operate at widely different conceptual levels. For example, article 12 of the UN Declaration of Human Rights[240] and Privacy by Design[241] contain high-level privacy principles. Fair information practice principles[242] and contextual integrity[243] provide mid-level guidance. Privacy impact assessments,[244] k-anonymity,[245] and traditional statistical disclosure limitation techniques[246] are examples of applied methods for enhancing confidentiality. Proposals such as differential privacy[247] incorporate formal mathematical frameworks for privacy. Finally, some proposals at the individual level include privacy policy "nutrition labels"[248] and personal data stores.[249]

---

237 5 U.S.C. §§ 552a(b), (a)(4).

238 *See* Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, § 512(b), 116 Stat. 2899, 2966 (2002) (codified at 44 U.S.C. § 3501 note (2013)).

239 *See* Proposed Rule, Improve Tracking of Workplace Injuries and Illnesses, 78 Fed. Reg. 67254, 67257–60 (Nov. 8, 2013).

240 Universal Declaration of Human Rights art. 12, G.A. Res. 217 (III) A, U.N. Doc. A/RES/217(III) (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.").

241 Ann Cavoukian, Privacy by Design 1 (2009), https://www.privacybydesign.ca/content/uploads/2009/01/privacybydesign.pdf.

242 *See* ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, THE OECD PRIVACY FRAMEWORK (2013), 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 (May 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 229.

243 *See* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119 (2004).

244 *See* DAVID WRIGHT & PAUL DE HERT, PRIVACY IMPACT ASSESSMENT (2012).

245 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 225.

246 *See* Gregory J. Matthews, *Data Confidentiality: A Review of Methods for Statistical Disclosure Limitation and Methods for Assessing Privacy*, 5 STATISTICAL SURVEYS 1 (2011).

247 *See* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMUNICATIONS OF THE ACM 86 (2011).

248 *See* Patrick Gage Kelley et al., *A "Nutrition Label" for Privacy*, 5 SYMP. ON USABLE PRIVACY & SECURITY, Article No. 4 (2009).

249 *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.

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 privacy enhancing methods and systems, or for evaluating the privacy considerations related to a particular data release case. As mentioned above, 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,[250] information privacy literature describes many controls, threats, vulnerabilities, and measures of utility, but no catalog or ready categorizations exist for privacy-related factors.

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. 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,[251] 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.[252] 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.[253]

---

250 This paper 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, Security and Privacy Controls for Federal Information Systems and Organizations, Special Publication 800-53 (2013).

251 Data perturbation refers to the masking 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 17.

252 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).

253 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).

- Privacy *vulnerabilities* are defined as characteristics that increase the likelihood that threats will be realized.[254] 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 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. 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.

We believe this article is the first to adopt this categorization explicitly and to use the specific definitions above.[255] However, elements of this categorization are closely related not only to information security definitions, as mentioned, but also to a line of prior work in several other fields.[256]

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 science and in records management,[257] documents the information objects, actors,

254 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.

255 This may be considered a formalization of the framework we and our collaborators sketch in prior work. *See, e.g.*, Salil Vadhan et al., Comments to the Department of Health and Human Services Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket No. HHS-OPHS–2011–0005 (Oct. 26, 2011), http://privacytools.seas.harvard.edu/files/privacytools/files/commonruleanprm.pdf; 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 138 (explicitly characterizing the primary privacy threat models for releases of official statistics); Dwork, *supra* note 28 (defining differential privacy in terms of a specific combination of threat model, vulnerability characterization, and choice of control); Wu, *supra* note 24 (providing an overview and detailed analysis of threat models used in the privacy context); NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, PRIVACY RISK MANAGEMENT FOR FEDERAL INFORMATION SYSTEMS, DRAFT OF NISTIR 8062 (May 2015) (proposing draft guidance to characterize privacy in terms of controls, threats, and risks, using somewhat narrower definitions than those we adopt); 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).

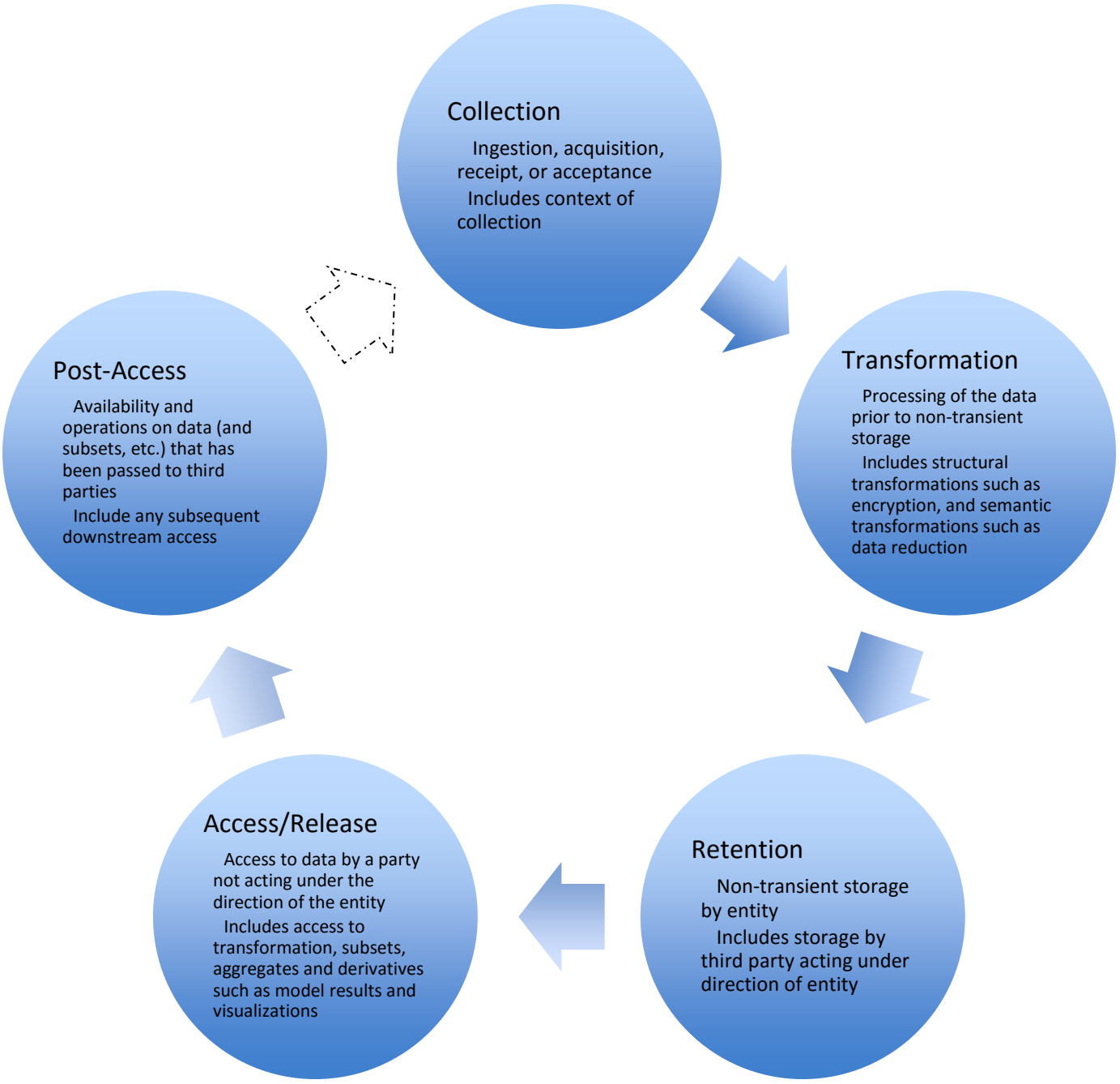
257 Our proposed framework incorporates a partial "lifecycle" model that focuses on the stages of activity associated with government data releases. Lifecycle models have been used in biology for at least a hundred years. They have been applied to processes in many fields, such as project management 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

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 provide no 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.

Figure 1. A lifecycle model for government data releases, based on use cases in Part II.

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.



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.

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.

Some sort of classification of controls is clearly needed to provide guidance. Of information security standards, FISMA[258] and the implementing guidelines[259] from NIST provide the most systematic and extensive classifications of controls. Moreover, with NIST's latest draft guidelines,[260] 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.[261]

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 system; data subjects, controllers, or recipients generally; or the public at large about privacy practices and risks;

---

[258] Federal Information Security Management Act, 44 U.S.C. §§ 3541-3549.

[259] See, e.g., NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (NIST), STANDARDS FOR SECURITY CATEGORIZATION OF FEDERAL INFORMATION AND INFORMATION SYSTEMS, FIPS Publication 199 (2004); NIST, MINIMUM SECURITY REQUIREMENTS FOR FEDERAL INFORMATION AND INFORMATION SYSTEMS, FIPS Publication 200 (2006); NIST, GUIDE FOR DEVELOPING SECURITY PLANS FOR FEDERAL INFORMATION SYSTEMS, Special Publication 800-18 (2006).

[260] NIST, PRIVACY RISK MANAGEMENT FOR FEDERAL INFORMATION SYSTEMS, Internal Report 8062 (Draft) (May 2005), http://csrc.nist.gov/publications/drafts/nistir-8062/nistir_8062_draft.pdf.

[261] See NIST, SECURITY AND PRIVACY CONTROLS FOR FEDERAL INFORMATION SYSTEMS AND ORGANIZATIONS, Special Publication 800-53 (2013).

- *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.[262] Consent mechanisms are evolving, and there is movement in some areas towards more portable and broader consent for certain uses of information, such as research uses.[263] 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.[264] 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.[265] 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 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

262 *See, e.g.*, Jeff Jonas & Jim Harper, *Open Government: The Privacy Imperative*, *in* OPEN GOVERNMENT (Daniel Lathrop & Laurel Ruma eds., 2010).
263 *See* Effy Vayena et al., *Caught in the Web: Informed Consent for Online Health Research*, 5 SCI. TRANSLATIONAL MED. 173fs6 (2013).
264 *See* Lorrie Faith Cranor, *Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice*, 10 J. ON TELECOMMUNICATIONS & 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).

"nutrition labels."[266] 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,[267] though participation can be incentivized by offering payments to individuals who agree to share their information.[268] 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.[269]

In addition to notice and consent, agencies often seek to provide privacy protection at the acceptance stage by implementing several other fair information practice principles: collection limitation, data minimization, and purpose specification in the design of a data collection program.[270] For instance, governments may prohibit the collection of personal information except for specific, limited purposes.[271] 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.[272] 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.[273] 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.[274] All federal executive agencies are required to conduct privacy impact assessments for information technology systems containing personally identifiable

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 248; Aza Raskin & Arun Ranganathan, *Privacy: A Pictographic Approach*, W3C Workshop on Privacy for Advanced Web APIs (2010); W3C, The Platform for Privacy Preferences 1.0 (P3P1.0) Specification (Apr. 16, 2002), http://www.w3.org/TR/P3P.

267 O'HARA, *supra* note 265, at 49–50.

268 Bart van der Sloot, *On the Fabrication of Sausages, or of Open Data and Private Data*, EJOURNAL EDEMOCRACY & OPEN GOV'T 136 (2011).

269 O'HARA, *supra* note 265, at 49–50.

270 *See* U.S. DEP'T OF HEALTH, EDUC., & WELFARE, *supra* note 229.

271 *See* Scassa, *supra* note 20.

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*, Proceedings of the CEDEM13 Conference (2010); Ugo Pagallo & Eleonora 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).

information.[275] 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.[276] Note, however, that such assessments do not generally include documenting specific privacy threats or vulnerabilities. Section III.C details this shortcoming.

2. Privacy controls at the transformation stage

Transformation of data includes 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.[277] 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.[278] 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 direct and quasi-identifiers.[279] Fields are typically redacted, according to varying standards such as the HIPAA Privacy Rule safe harbor de-identification standard,[280] 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.[281] 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.[282]

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

275 *See* OFFICE OF MGMT. & BUDGET, *supra* note 218.
276 *See id.*
277 For a detailed description of public- and private-key encryption standards for federal government information systems, see NIST, SECURITY REQUIREMENTS FOR CRYPTOGRAPHIC MODULES, FIPS 140-2 (2001), http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf.
278 *See* Hugo Krawczyk, *Secret Sharing Made Short,* PROCEEDINGS OF THE 13TH ANNUAL INT'L CRYPTOLOGY CONFERENCE (1993).
279 *See* Thomas P. Keenan, *Are They Making Our Privates Public? Emerging Risks of Governmental Open Data Initiatives*, *in* PRIVACY AND 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 17.
282 *See* Alan F. Karr & Jerome P. Reiter, *Analytical Frameworks for Data Release: A Statistical View*, in CONFIDENTIALITY AND DATA ACCESS IN THE USE OF BIG DATA: THEORY AND PRACTICAL APPROACHES (2014).

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.[283]

Another approach is to generate synthetic data from a statistical model that has been developed using the original data set. 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.[284] They are, however, of 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 strength of 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,[285] 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.[286] 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.[287] 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.[288] 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.[289]

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

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 STATISTICS 115 (1994); Donald B. Rubin, *Discussion of Statistical Disclosure Limitation*, 9 J. OFFICIAL STATISTICS 461 (1993).

285 Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMUNICATIONS OF THE ACM 86 (2011).

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 ENVIRONMENT & PLANNING A 1101 (2007).

287 *See* Goodin, *supra* note 89.

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.

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.[290] For example, for some categories of confidential data, industry standards may require encryption,[291] or laws may require encryption where reasonable.[292] 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.[293] Many states require personal information maintained by state agencies or businesses to be destroyed when the data are no longer needed.[294] 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.[295]

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.[296] 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.[297]

Privacy dashboards and personal data stores are tools which individuals can use to express detailed permissions regarding retention and uses of their data. An individual can use a web-based privacy dashboard to grant granular access permissions to her data only to select parties or for select uses.[298] 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.[299] 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.[300] Legal interventions include statutory breach reporting requirements, which require organizations to notify individuals and enforcement bodies in the event of a data security breach. [301] Examples of technical measures include federated databases, for enabling

290 *See, e.g.*, U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, CONTROLS (final), *supra* note 229.
291 *See, e.g.*, SECURITY STANDARDS COUNCIL, PAYMENT CARD INDUSTRY DATA SECURITY STANDARD: REQUIREMENTS AND SECURITY ASSESSMENT PROCEDURES (Apr. 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.*, Family Educational Rights and Privacy Act, 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) (2013); van der Sloot, *supra* note 268, at 144.
297 *See* Scassa, *supra* note 20.
298 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 197.
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).

controlled queries across databases maintained by different organizations,[302] computable policies, for automating the enforcement of privacy policies,[303] and secret sharing and other techniques for managing keys for encrypted systems.[304]

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 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,[305] while statistical agencies have required strong confidentiality protections for their data.[306] Governments are increasingly pressured to make information available under a presumption of openness,[307] and commentators have suggested that expert panels, including a broad range of stakeholders, be involved in developing policies for making release decisions.[308] 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.[309]

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.

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

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.
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 Communications of the ACM 612 (1979).
305 *See* Conley et al., *supra* note 20, at 778.
306 *See, e.g.*, Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, 116 Stat. 2899, 2962 (2002) (codified at 44 U.S.C. § 3501 note (2013).
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.

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.[310] 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.[311] 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 dataset 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.[312]

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. 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 these forms of data sharing have even been shown to be compatible with a strong new privacy guarantee known as differential privacy.[313] Differential privacy provides a framework for measuring and reducing the risk of disclosing privacy-sensitive information about individuals when analyzing and sharing data.[314] 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.[315]

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.[316] 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,

---

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.*, Census Bureau, Census RDC Research Proposal Guidelines (2015), http://www.census.gov/ces/pdf/Research_Proposal_Guidelines.pdf.

312 U.S. Census Bureau, *supra* note 153.

313 For the foundations of differential privacy, *see* Irit Dinur & Kobbi Nissim, *Revealing Information while Preserving Privacy*, 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*, Proceedings of the 24th Annual International Cryptology Conference 528 (2004); Avrim Blum, Cynthia Dwork, Frank McSherry, & Kobbi Nissim, *Practical Privacy: The SuLQ Framework*, 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*, Proceedings of the 3rd Conference on Theory of Cryptography 265 (2006).

314 *See* Dwork, *Differential Privacy*, *supra* note 28.

315 *See* Dwork, *supra* note 247.

316 *See* Yehuda Lindell & Benny Pinkas, *Secure Multiparty Computation for Privacy-preserving Data Mining*, 1 J. PRIVACY & CONFIDENTIALITY 59 (2009); *See* Alan F. Karr et al., *Secure Regression on Distributed Databases*, 14 JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS 263 (2005).

and allows computations across parties that do not fully trust each other.[317] In theory, it can be combined with the interactive mechanisms and privacy aware computational methods previously mentioned.[318]

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 an encryption method being developed to enable computations to be performed on encrypted data without decrypting the data first and exposing it to attack.[319]

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.[320] 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.[321]

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.

Transparency and accountability for misuse are essential to achieving an optimal balance of social benefit and individual privacy protection.[322] In the face of ubiquitous data collection practices, individuals find it difficult to effectively withhold consent because the playing field is uneven,[323] making accountability for misuse of increasing importance. In addition, data collectors and individual

317 *See* sources cited *supra* note 316.
318 *See, e.g.*, Amos Beimel, Kobbi Nissim, & Eran Omri, *Distributed Private Data Analysis: Simultaneously Solving* How *and* What, PROCEEDINGS OF THE 28TH ANNUAL INT'L CRYPTOLOGY CONFERENCE (CRYPTO) 451 (2008) (exploring the combination of secure multiparty computation and differential privacy).
319 *See* Craig Gentry, *Fully Homomorphic Encryption Using Ideal Lattices*, PROCEEDINGS OF THE ACM SYMPOSIUM ON THEORY OF COMPUTING (STOC) 169 (2009).
320 *See, e.g.*, O'HARA, *supra* note 265.
321 *See* Conley et al., *supra* note 20, at 843–44.
322 Daniel J. Weitzner et al., *Information Accountability*, 51 COMMUNICATIONS OF THE ACM 82 (2008).
323 *See* Paul M. Schwartz & Daniel Solove, *Notice and Choice: Implications for Digital Marketing to Youth*, Second NPLAN/BMSG Meeting on Digital Media and Marketing to Children (2009).

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.[324] Non-governmental organizations, privacy commissioners, and the public should be able to monitor government releases of data and speak out about privacy violations.[325] 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,[326] 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 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,[327] though there have been recent proposals to automate the generation of custom data use agreements.[328]

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.[329] 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

---

324 *See, e.g.*, Molinari & Marsh, *supra* note 274, at 313–14; van der Sloot, *supra* note 268, at 149.
325 *See, e.g.*, Keenan, *supra* note 279.
326 *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 data sets).
327 O'HARA, *supra* note 265, at 109.
328 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 a modular data use agreement generator.
329 *See* Jonas & Harper, *supra* note 262.

and vulnerabilities at each stage subject to maintaining the desired uses and expected benefits of the data.

Table 1. Example categorization of privacy controls and interventions.

	Procedural	Economic	Educational	Legal	Technical
Collection/ Acceptance	Collection limitation; Data minimization; Data protection officer; Institutional review boards; Notice and consent procedures; Purpose specification; Privacy impact assessments;	Collection fees; Markets for personal data; Property rights assignment	Consent education; Transparency; Notice; Nutrition labels; Public education; Privacy icons	Data minimization; Notice and consent; Purpose specification	Computable policy
Transformation	Process for correction		Metadata; Transparency	Right to correct or amend; Safe harbor de-identification standards	Aggregate statistics; Computable policy; Contingency tables; Data visualizations; Differentially private data summaries; Redaction; SDL techniques; Synthetic data
Retention	Audits; Controlled backups; Purpose specification; Security		Data asset registers; Notice; Transparency	Breach reporting requirements; Data retention and destruction requirements;	Computable policy; Encryption; Key management (and Secret sharing);

	assessments; Tethering			Integrity and accuracy requirements	Federated databases; Personal data stores
Access/Release	Access controls; Consent; Expert panels; Individual privacy settings; Presumption of openness vs. privacy; Purpose specification; Registration; Restrictions on use by data controller; Risk assessments	Access/Use Fees (for data controller or subjects); Property rights assignment	Data asset registers; Notice; Transparency	Integrity and accuracy requirements; Data use agreements (contract with data recipient)/ Terms of service	Authentication; Computable policy; Differential privacy; Encryption (incl. Functional; Homomorphic); Interactive query systems; Secure multiparty computation
Post-Access (Audit, Review)	Audit procedures; Ethical codes; Tethering;	Fines	Privacy dashboard; Transparency	Civil and criminal penalties; Data use agreements/ Terms of service; Private right of action	Computable policy; Immutable audit logs; Personal data stores

C. IDENTIFYING INFORMATION USES, THREATS, AND VULNERABILITIES

Assessment and treatment 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. [330] 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 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;

330 Vadhan et al., *supra* note 25.

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, and the organizations releasing data, and analysts who seek to use it, 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 methodology 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 only aggregate-level data to enable statistical analyses at the population level, but not learning about individual respondents in the data.[331] 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[332] or political donors,[333] 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.[334] 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.[335] As required by law, the city's police department released at least 2.1 million records including the date, time, and location of automobiles throughout the city.[336] 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 data about them 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.[337] The intended law enforcement use of the data and the public safety

331 *See, e.g.*, Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, tit. V, § 512 (b)(1), 116 Stat. 2899, 2966 (2002) (codified at 44 U.S.C. § 3501 note (2013)).
332 Fitz-Gibbon, *supra* note 80.
333 *See* ProtectMarriage.com v. Bowen, 752 F.3d 827, 835 (9th Cir. 2014).
334 *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.
335 *See* Minn. Stat. § 13.03 (2012).
336 *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.
337 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).

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 release was not well-matched to the intent and expectations 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 cause harm to 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 factors related to 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 such threats is a useful first step in estimating the extent and severity 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, if, for example, the only conceivable harm is 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.[338] 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.[339] 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., mistaken release of data or software defects), as such events are known to be a significant risk in data management.[340] Privacy threats of environmental origin (e.g., due to 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.[341]

Note that 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

338 For a general detailed and thoughtful discussion of threat models in the privacy context, see Wu, *supra* note 24.
339 *See* U.S. NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, CONTROLS (draft), *supra* note 191.
340 *See, e.g.*, Stephen Ohlemacher, *Census Bureau Admits Privacy Breach*, USA TODAY (Mar. 7, 2007), http://usatoday30.usatoday.com/news/washington/2007-03-07-1535966293_x.htm (reporting that the Census Bureau inadvertently posted personal information publicly while testing new software).
341 *See* WILLENBORG & DE WAAL, *supra* note 138.

labeled "quasi-identifiers," whereas differential privacy assumes no limits on auxiliary adversary knowledge. Note, 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, 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."[342] 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.[343] 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 a vulnerable group such as stereotyping, price discrimination against vulnerable groups, 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.[344] 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, 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,

342 *See* Michal Kosinski et al., *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 5802 (2013).

343 *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).

344 *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).

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.[345] In the late 1990s, Latanya Sweeney identified the record of Massachusetts Governor William Weld in an anonymized medical claims dataset by comparing sex, ZIP code, and date of birth with publicly available voter registration records.[346] These three seemingly innocuous pieces of information uniquely identify well over 50% of the U.S. population.[347] To mitigate the risk of identity disclosure, organizations typically make efforts to "anonymize" 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.[348]

It is now well-understood, however, that stripping direct and quasi-identifiers provides very weak privacy protections, as it is often quite easy to reidentify individuals in data that have been treated in

345 *See* Alan F. Karr & Jerome P. Reiter, *Using Statistics to Protect Privacy*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD (Julia Lane et al. eds., 2014).
346 *See* Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J. L., MED., & ETHICS 98; Latanya Sweeney, *Uniqueness of Simple Demographics in the US Population*, Data Privacy Lab Technical Report (2000).
347 *See* sources cited *supra* note 346.
348 *See, e.g.*, 45 C.F.R. § 164.514(b) (2014); 34 C.F.R. § 99.31(b) (2014).

this way.[349] It has been shown more generally that it takes very little information to uniquely identify an individual.[350] 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.[351] 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.[352] There have been numerous other examples where this phenomenon has been exploited for reidentification,[353] and disclosure risks continue to grow as information about individuals is increasingly made available through publicly accessible government and commercial databases.[354]

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.[355] 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.[356] 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.[357] 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. It has been shown, for example, 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.[358] Researchers have discovered attribute disclosure risks in recommendation systems such as Amazon's system for

349 *See, e.g.*, Ohm, *supra* note 23.
350 *See* de Montjoye et al., *supra* note 226; Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 NATURE SCI. REP. 1376 (2013).
351 *See* sources at *supra* note 350.
352 *See* Narayanan & Shmatikov, *supra* note 225.
353 *See, e.g.*, Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES Aug. 9, 2006.
354 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.
355 *See, e.g.*, Josep Domingo-Ferrer & Vicenç Torra, *Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage*, 13 STATISTICS & COMPUTING 343 (2003).
356 *See* Sweeney, *supra* note 225.
357 *See* Karr & Reiter, *supra* note 345.
358 *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*, LOS ANGELES TIMES (Aug. 29, 2008), http://articles.latimes.com/2008/aug/29/local/me-dna29.

providing product suggestions based on aggregated consumer behavior.[359] As another 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 than one thousand individuals by querying the system and, furthermore, demonstrated that it was possible to link the records to identifiable people.[360]

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.[361] 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.[362] 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.[363]

D. DESIGNING DATA RELEASES BY ALIGNING USE, THREATS, AND VULNERABILITIES WITH CONTROLS

In this Part, we have described the elements of a framework for 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 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

359 *See* Joseph A. Calandrino et al., *"You Might Also Like:" Privacy Risks of Collaborative Filtering*, Proceedings of the IEEE Symposium on Security and Privacy 231 (2011).
360 *See* Ziv, *supra* note 21.
361 *See* NAT'L RESEARCH COUNCIL, PUTTING PEOPLE ON THE MAP *supra* note 18.
362 *See* de Montjoye et al., *supra* note 350.
363 *See* de Montjoye et al., *supra* note 226.

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 often to be 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 analytical uses required, it may nevertheless be useful to consider the analytical characteristics of the intended uses of the data, including the desired form of the analytical output; the goal of the analysis; the utility, loss, or quality measure; and the analysis methodology.[364] In addition, the compatibility of controls should be considered in light of the proposed analytical 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 analytical 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.AC, 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,[365] the following data characteristics are particularly relevant to an analysis of vulnerabilities in a data release and the selection of appropriate controls:

- 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;

364 *See* Alexandra Wood et al., *Integrating Approaches to Privacy Across the Research Lifecycle: Long-term Longitudinal Studies*, (Working Paper 2014), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2469848.
365 *See id.*

- 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.[366] Rich, messy data, such as information from social networks, can contain unanticipated information in the structure of the data itself that creates vulnerabilities, as the identifiability of the data will likely remain high after standard redaction controls have been applied.[367]

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 harm from an uncontrolled release can help guide the selection of an appropriate set of privacy controls.

Figure 2 below provides a partial conceptualization of the relationship between the threats and vulnerabilities associated with a given set of data and the suitability of selected procedural and legal controls implemented at the collection and release stages. Note that, 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. 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 practice.

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. In between these two end

366 *See, e.g.*, Narayanan & Shmatikov, *supra* note 225.

367 *See, e.g.*, Lars Backstrom et al., *Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography*, Proceedings of the 16th Int'l Conference on World Wide Web 181 (2007); Jasmine Novak et al., *Anti-Aliasing on the Web*, Proceedings of the 13th Int'l Conference on World Wide Web 30 (2004).

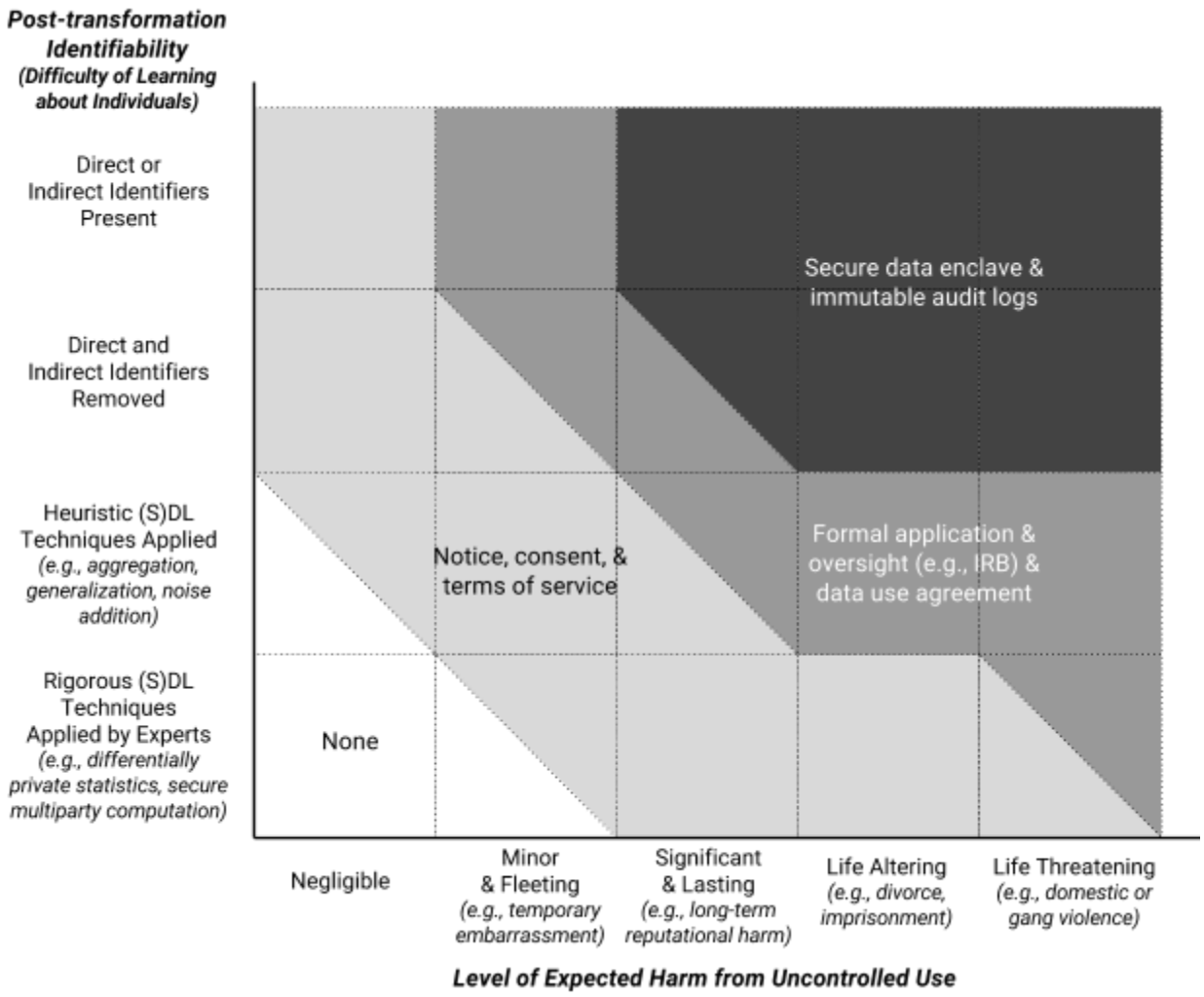
points fall examples of minor and temporary harms, significant and lasting harms, and life altering harms that fall short of being life threatening.

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, there are data sets containing direct or indirect identifiers, such as names, addresses, and dates of birth. At the other end, there 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 encryption. In between, there are examples such as data sets 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.

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.



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 releasing data from this category. An example of data described by this category are national and state test scores released as custom aggregate statistics by the Department of Education's National Center for Education Statistics under terms of service prohibiting re-identification and linking of the data, among other restrictions.[368] For some data within this category, such as data collected from human subjects for research conducted with federal funding, approval from an institutional review board and a data use agreement may be required, as reflected by the medium gray regions.[369]

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, and data use agreements are

[368] National Center for Education Statistics, NAEP Data Explorer (last visited Aug. 10, 2015), http://nces.ed.gov/nationsreportcard/naepdata.
[369] *See* the Federal Policy for Human Subjects Research, 45 C.F.R. pt. 46.

used to limit data use and set 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.[370]

For highly identifiable and harmful data, represented by the darkest gray region, access is 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,[371] maintained by statistical agencies in identifiable form and therefore protected by CIPSEA.[372]

In some cases, the practices represented by this diagram deviate from 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 deidentification standard or its limited data set standard for deidentification. If redacted according to the safe harbor standard, the data can be released without any restriction, and, if redacted according to the limited data set 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.[373] In light of evolving best practices, an agency may decide not to adopt this standard but to require privacy controls, such as application and oversight procedures, that are more restrictive than the law 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.

As indicated by the diagram, 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 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 data set, including direct and indirect identifiers, could be instructed to submit an application to an institutional review board or

---

[370] *See* FERPA, 34 C.F.R. §§ 99.31(a)(6)(iii)(C), 99.35(a)(3); HIPAA Privacy Rule, 45 C.F.R. §§ 164.512(i)(1)(i), 45 C.F.R. 164.514(2).

[371] *See* U.S. Census Bureau Data Stewardship Executive Policy Committee, Policy on Respondent Identification and Sensitive Topics in Dependent Interviewing, Policy DS-16 (Dec. 9, 2014), http://www2.census.gov/foia/ds_policies/ds016.pdf.

[372] For a discussion of the use of formal screening processes and secure data enclaves for accessing statistical information, see *supra* section II.A.3.

[373] *See* discussion *supra* Section III.C.3.

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 created at the collection stage if the data collection process could be observed by an adversary; data that are retained long-term 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 should be considered to be representative of many of the types of issues that arise in government data releases rather than issues specific 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.[374] 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

[374] Improve Tracking of Workplace Injuries and Illnesses; Proposed Rule, 78 Fed. Reg. 67254 (Nov. 8, 2013).

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.[375] Furthermore, the costs of occupational injury in the United States are at minimum tens of billions of dollars annually.[376] 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.[377] 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 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.[378] The proposed rule would expand the scope of collection to include all incident-specific injury and illness records currently maintained by these companies.[379] 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.[380] 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.[381] 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.[382] 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 privacy risks and potential harms for the individuals involved. For these reasons, the

375 *See* A. Mitchell Polinsky & Steven Shavell, *The Economic Theory of Public Enforcement of Law*, 38 J. ECON. LITERATURE 45 (2000).

376 *See* J. Paul Leigh, *Economic Burden of Occupational Injury and Illness in the United States*, 89 MILBANK Q. 728, 728-29 (2011).

377 *See id.*

378 U.S. Dep't of Labor, Improve Tracking of Workplace Injuries and Illnesses; Proposed Rule, 78 Fed. Reg. 67254, 67263 (proposed Nov. 8, 2013).

379 *Id.*

380 *Id.* at 67259-60.

381 *See, e.g.*, Sweeney, *supra* note 21.

382 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).

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,[383] 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 rulemaking calls for the public release of all workplace illness and injury records collected by OSHA, and it does not require OSHA to transform the data in any way prior to release. For instance, it does 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.[384] 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

383 Federal Information Security Management Act of 2002, 44 U.S.C. §§ 3541–49 (2013).

384 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).

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 it 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.[385] 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 injury or illness and where and how it occurred.[386] 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.[387] 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.[388] 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.[389] 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.

385 U.S. Dep't of Labor, Improve Tracking of Workplace Injuries and Illnesses, 78 Fed. Reg. 67254, 67263 (proposed Nov. 8, 2013).
386 *Id.* at 67259–60.
387 *See id.* at 67263. A final rule is anticipated to be published in August 2015.
388 *See* Occupational Safety and Health Administration, Follow-on Mockup to Proposed Web-Based Mechanism for OSHA's Injury/Illness Data Collection: Public Access to Data (Apr. 22, 2013), https://www.osha.gov/recordkeeping/LDCsys-rulemaking-Search.pdf.
389 *See id.*

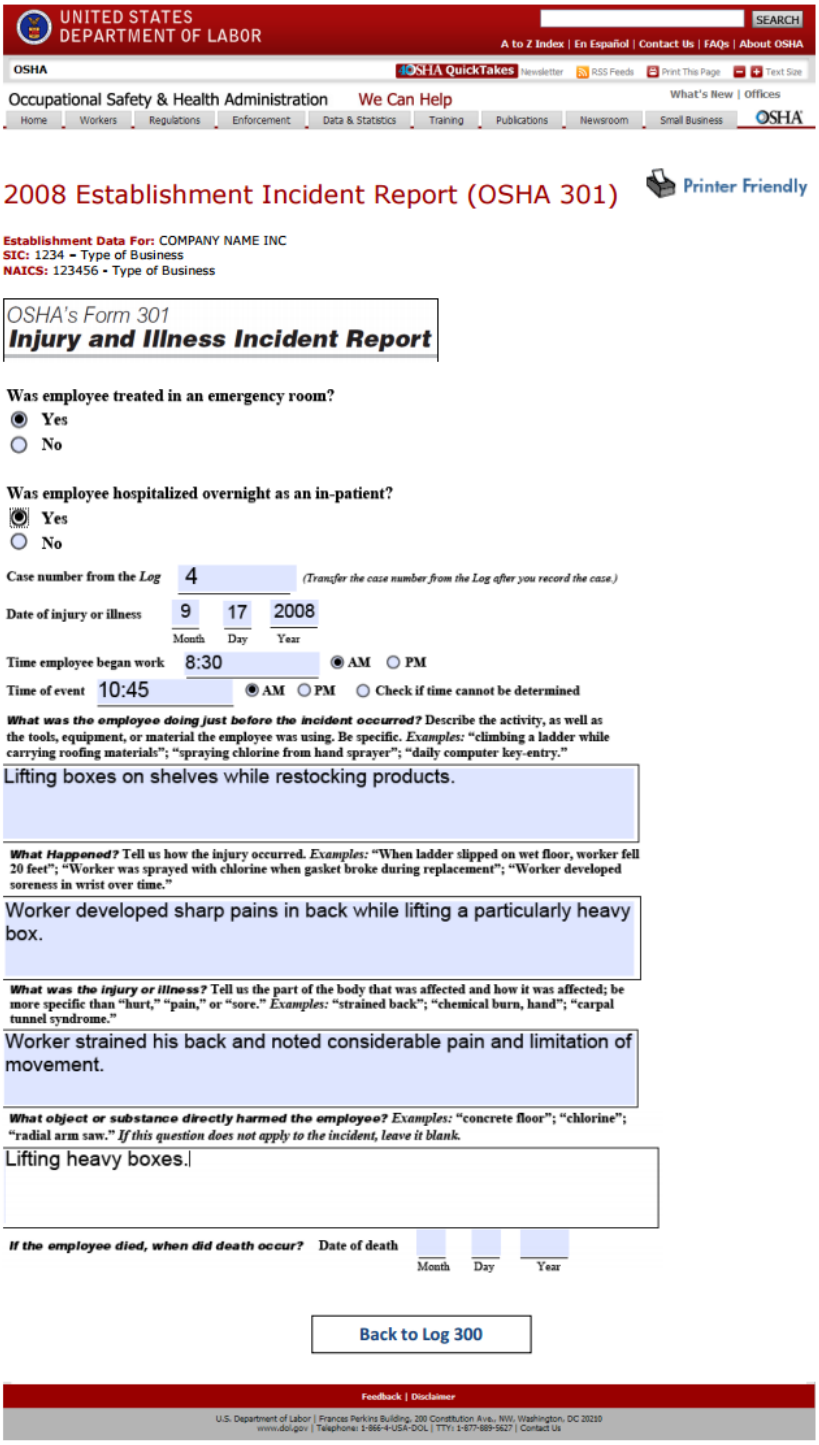


Figure 3. OSHA's mockup of proposed web display of workplace injury and illness reports.[390]

390 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.

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,[391] 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.[392]

The uniqueness of the individual record 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 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."[393] 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.[394]

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 are required to be removed.[395] 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

391 5 U.S.C. § 552a(a)(5) (2013).

392 Legal scholars have also raised a number of concerns regarding the privacy protections for individuals in FOIA. *See, e.g.*, Bloom, *supra* note 20; 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).

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 130 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.

395 78 Fed. Reg. 67254, 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).

reduces the utility of the released data for scientific and policy analysis, and indicates 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 by the public.[396] 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.[397] 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 of their personal information.

6. Aligning uses, threats, and vulnerabilities with privacy 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 information. Despite these uncertainties, the rulemaking does not propose additional privacy controls, such as requiring the release of only aggregate records, the generalizing of free-form text fields as categorical values, or the use of other more advanced techniques to transform the data to provide stronger privacy protections. It 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 align with 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

396 For example, 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.).

397 *See* Weitzner et al., *supra* note 322.

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.[398] 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,[399] 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 approaches. 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 one 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.[400] 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 to help employers identify information within their workplace injury records that poses a disclosure risk could offer helpful guidance, especially if OSHA expects that its recordkeeping forms will continue 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.

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,[401] and the City of Seattle has released over 300 datasets.[402] 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),[403] customizable maps and charts, and feeds to an API that enables apps to

398 *See, e.g.*, Dwork, et al, *supra* note 313.
399 Julia Lane & Stephanie Shipp, *Using a Remote Access Data Enclave for Data Dissemination*, 2 INT'L J. DIGITAL CURATION 128 (2007).
400 *See* National Research Council reports cited at *supra* note 18.
401 *See Data Boston*, CITY OF BOSTON, https://data.cityofboston.gov (last visited July 15, 2015).
402 *See Results Matching Type of Datasets*, DATA.SEATTLE.GOV, https://data.seattle.gov/browse?limitTo=datasets (last visited July 15, 2015).
403 *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 data.[404] 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.

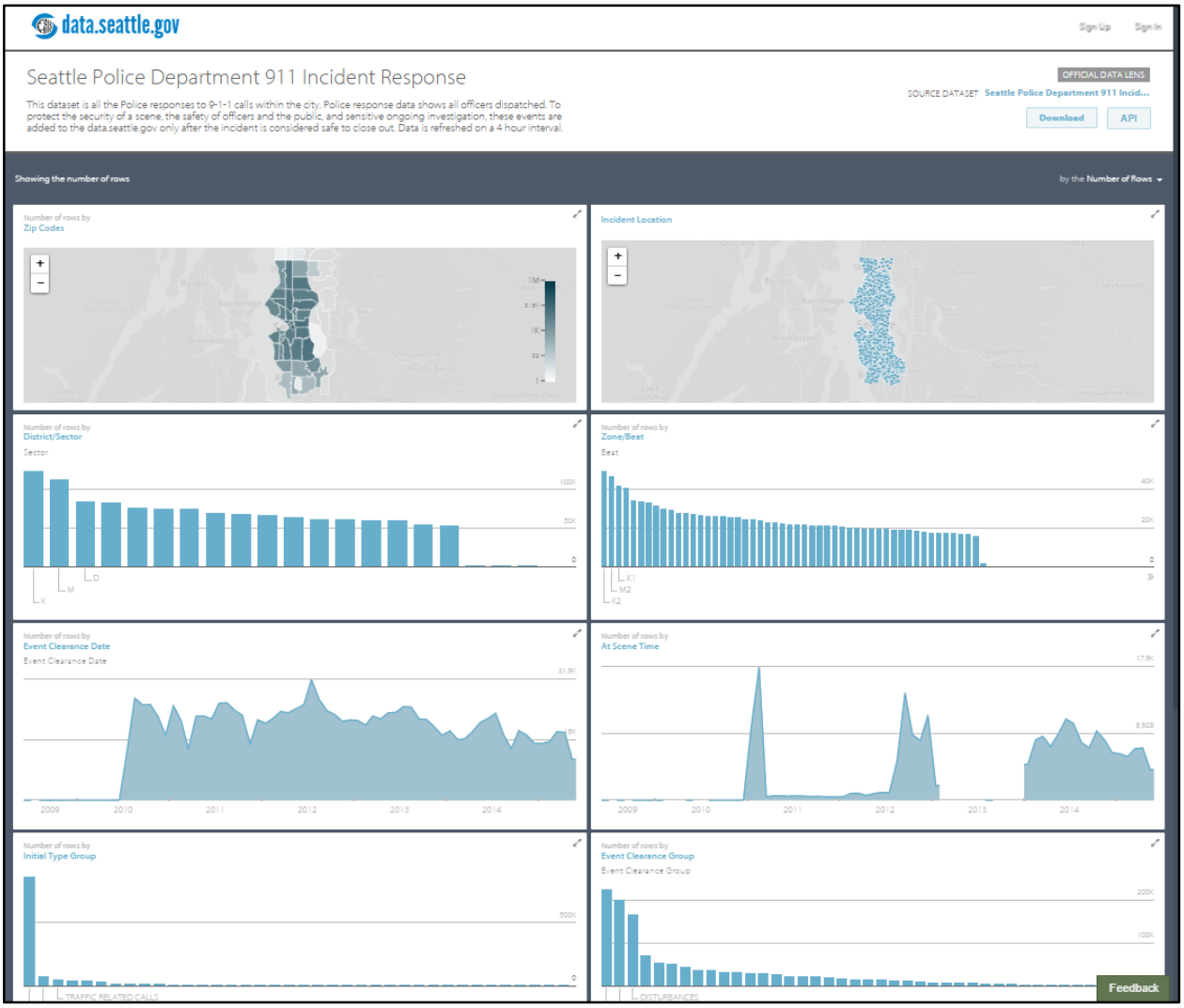


Figure 4. Screenshot of display of "data lens" visualization of police department 911 incident response data from the City of Seattle open data portal.

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

404 *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).

and provide services to city residents.[405] 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.[406] 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 may be particularly vulnerable to reidentification or learning 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 a collection limitation principle, inappropriate. In addition, the city data managers should be transparent to the public about the scope of data collection, 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

[405] *See Data Boston: Results matching type of Datasets*, CITY OF BOSTON.GOV, 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).

[406] *See, e.g.*, *Street Bump*, BOSTON MAYOR'S OFFICE OF NEW URBAN MECHANICS, http://www.streetbump.org (last visited Aug. 17, 2015) (describing an app that uses a smartphone's built-in sensors to detect potholes that volunteers encounter while driving throughout the city).

portal because some businesses are licensed under an owner's home address.[407] At this stage, the open data managers also remove or mask categories of sensitive information. For instance, the City of Boston removes all domestic violence and sexual assault cases from its crime incident data and generalizes descriptions for the remaining incidents using broad categories such as "drug charges."[408] 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.[409] 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.[410] 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.[411] 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 asthma in their home."[412] 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

---

407 *See* Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.

408 *See Crime Incident Reports*, DATA.CITYOFBOSTON.GOV, https://data.cityofboston.gov/Public-Safety/Crime-Incident-Reports/7cdf-6fgx (last visited May 26, 2015).

409 *See* Off-the-record interview with an open data manager for the City of Boston, Apr. 9, 2015.

410 *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).

411 *See Mayor's 24 Hour Hotline, Service Requests*, DATA.CITYOFBOSTON.GOV, https://data.cityofboston.gov/City-Services/Mayor-s-24-Hour-Hotline-Service-Requests/awu8-dc52 (last visited May 26, 2015).

412 *See Breathe Easy at Home*, CITYOFBOSTON.GOV, http://www.cityofboston.gov/isd/housing/bmc.asp (last visited May 26, 2015).

example, third party smartphone apps for tracking data points such as 911 calls to local police and fire departments,[413] 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 only information associated 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,[414] and is limited by regulations protecting certain categories of information, such as FERPA[415] for education records and state regulations for criminal offender record information.[416] 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,[417] 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.[418] 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.[419] They look to regulatory classifications of personally identifiable information from laws such as the HIPAA Privacy Rule; however, these laws are limited in scope and the lack of more comprehensive, formal guidance creates uncertainty. To

413 *See* DATA.SEATTLE.GOV, https://data.seattle.gov (last visited May 26, 2015).
414 Mayor of Boston, An Executive Order Relative to Open Data and Protected Data Sharing (Apr. 7, 2014), http://www.cityofboston.gov/news/Default.aspx?id=6589.
415 Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g (2013); 34 C.F.R. Part 99 (2013).
416 Mass. Gen. L. ch. 6, § 172 (2015).
417 Wash. Rev. Code § 42.56.001–.904 (2006).
418 *See* Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.
419 *See* Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.

address these concerns, the City of Seattle is developing formal governance procedures, requirements for reviewing and addressing disclosure risks in open data, and definitions for concepts like personally identifiable information.[420]

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.[421] For some of the 311 records, a field describing the reason for closing a case provides contact and contextual details about complaints, which can involve issues related to evictions and homelessness, medical conditions and disabilities, stalking incidents, and interpersonal relationship issues.[422] 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.[423] One record describes a domestic dispute involving child custody and visitation violations, restraining orders, and a relationship with a registered sex offender, as well as the phone number of the person who called the hotline.[424]

In other cases, a city intends to apply a privacy control but fails to implement it properly,[425] 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 the call.[426] 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.[427] 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.[428] 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 are not screening data adequately before release and selecting 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

420 *See* Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.
421 *See Mayor's 24 Hour Hotline Service Requests*, *supra* note 411.
422 *See Mayor's 24 Hour Hotline, Service Requests*, *supra* note 411.
423 *See id.*
424 *See id.*
425 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.
426 *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).
427 *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).
428 For a demonstration of this type of record linkage, *see, e.g.*, Sweeney, *supra* note 21.

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.[429] 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 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.[430] 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 repositories, and ones the cities should consider incorporating into their policies in order to mitigate disclosure risks at the post-access stage.[431]

429 *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).

430 *See id.*

431 For an example of standard terms of use implemented by one of the largest data repositories, *see, e.g.*, 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.

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.[432] Likewise, in transforming the data they rely on heuristics rather than formal standards to redact fields or collapse values into large categories.[433] 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.[434] 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 any 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 *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.

433 *See* sources cited at *supra* note 432.

434 *See, e.g.*, *South Boston, neighborhood: 2010 Census*" DATA.CITYOFBOSTON.GOV, https://data.cityofboston.gov/dataset/South-Boston-neighborhood-2010-Census/ybpb-72n5 (last visited May 26, 2015).

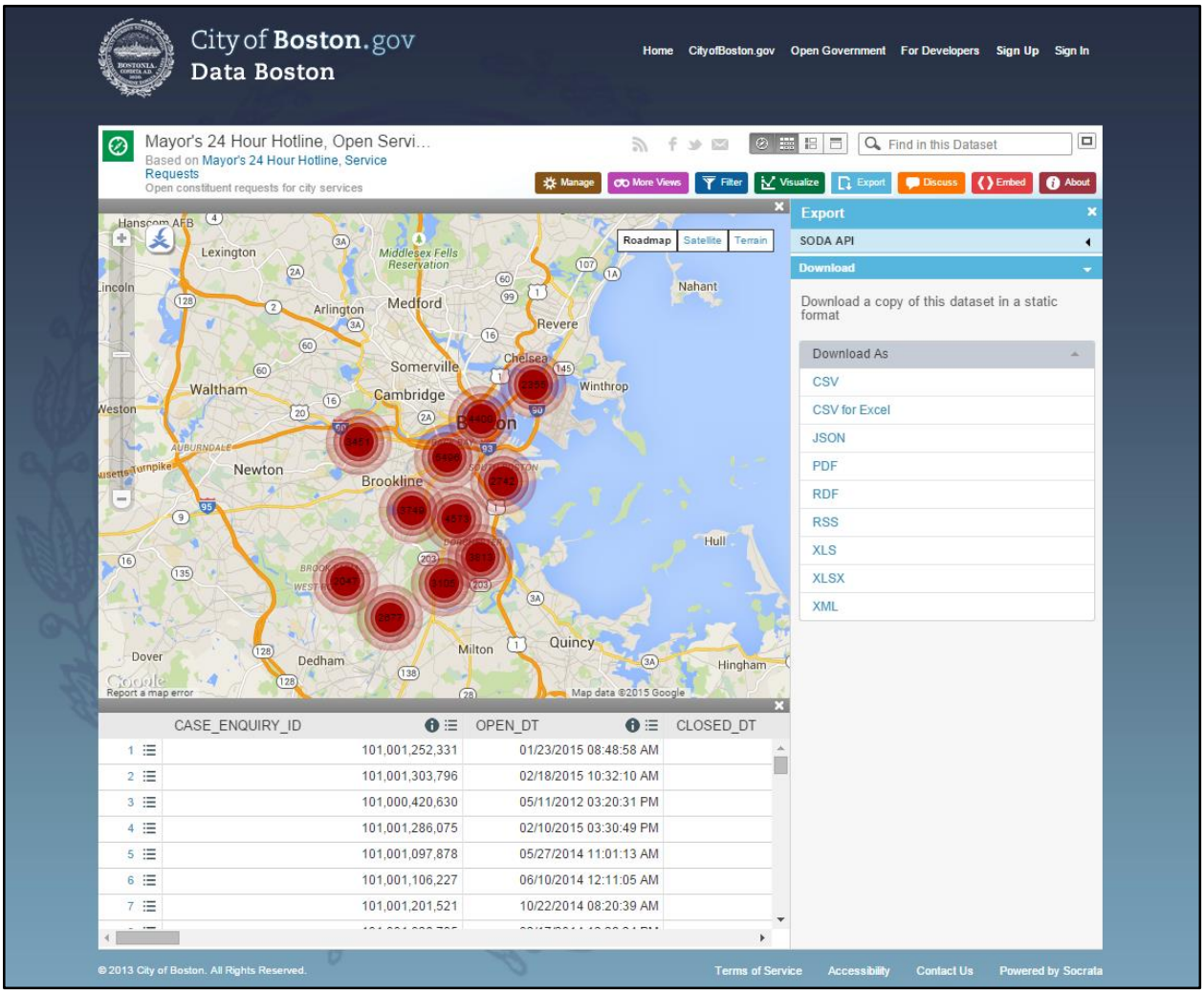


Figure 5. Visualization of 311 data in the City of Boston open data portal.

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 formal risk assessment model, as well as complementary data policy guidance for agencies and departments across the city to the follow.[435] 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.[436] In these efforts, the cities should 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 identifiable or sensitive information. Tiered access, as described more fully in the OSHA case in Section IV.A, allows for the implementation of

[435] *See* Off-the-record interview with an open data manager for the City of Boston, Apr. 9, 2015.
[436] *See* Off-the-record interview with open data managers for the City of Seattle, May 21, 2015.

privacy controls that are finely tuned to the intended uses, threats, and vulnerabilities relevant to the release. In this way, agencies can maximize public use of the data in ways that are socially beneficial 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 a wide range of uses. 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 has been 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 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 reinforce current concerns and outline a structure 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, are subject to very different assessments of privacy risk and selection 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 the inconsistencies in practice and the ineffectiveness of privacy safeguards adopted.

We find also that there are many gaps in the privacy controls used with government data releases. 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 based on the specific uses, threats, and vulnerabilities of the release.

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.