THE CENTER FOR
OPEN DATA ENTERPRISE

Discussion Draft:

Briefing Paper on Open Data and Privacy

Introduction

The federal government, like many governments around the world, now releases "open data" that can be used in a variety of ways. This important public resource has helped citizens investigate better value for college, fair housing, and safer medicines, and has a wide range of other potential public and economic benefits. Open data also helps government agencies themselves operate more efficiently, share information, and engage the citizens they serve. Under the U.S. government's Open Data Policy, all federal data should now be released as open data unless specific concerns, such as privacy or national security, prevent its open release.

Traditionally, "open government data" has been thought of as free, public data that anyone could use and republish. Now, the discussion is shifting to include data that may not be appropriate for wide, unfettered access, but can still be of use to non-government communities. For instance, data containing Personally Identifiable Information cannot be released widely, but there are certain circumstances that could allow for its use in restricted or de-identified forms. Various levels of sensitivity must be considered, leading to a range of potential levels of openness and methods for achieving that openness.[1]

Federal agencies must address the risks and sensitivities of making data available while at the same time maximizing its accessibility and use. To help find scalable solutions, the White House Office of Science and Technology Policy and the Center for Open Data Enterprise are co-hosting the 2016 Open Data Roundtables. These Roundtables bring together participants from federal agencies, academia, the private sector, and nonprofit organizations with technical, policy, ethical, and legal expertise. They are designed to:

- Identify open data case studies, learned lessons and best practices across the federal government;
- Strengthen a community of technical, legal, ethical, and policy experts in support of open data; and
- Support continuity and accelerate the progress of open data work.

[1] See, for example, the Open Data Institute's "Data Spectrum", https://theodi.org/data-spectrum.

This briefing paper presents an overview of the topic for the first Roundtable, which will cover privacy and open data. Privacy has become an urgent issue in data use: Most recently, in February 2016, President Obama recognized the need for clear guidelines by establishing the Federal Privacy Council.[2] This paper is meant to be used as a framing document, and is not by any means an exhaustive treatment of the subject. It is designed to support participants in their work and should not restrict them from covering other topics that they believe are central to the issue.

Open Data and Privacy

As more open government data has become available, data users in business, academia, and the nonprofit community have come up against a conundrum. Many datasets in health, education, housing, and other areas may have value when they are released with "microdata" that can be analyzed at the level of individual records. But releasing data at that level carries the risk of exposing personally identifiable information (PII) that could threaten individuals' privacy if it were released openly.

"Microdata" is *data released in its most granular, unaggregated form.*[3] For example, a list of the license plate and medallion numbers of every New York City taxi cab could be considered microdata. The key question is: *How can we maximize public access to and value from open granular information while protecting privacy?* To answer this question, data and privacy experts have explored issues such as:[4]

- What are the potential benefits of using unaggregated data (or microdata) for the public good?
- What are the risks of using these datasets if they contain or could lead to the discovery of personally identifiable information, and how can those risks be minimized?
- What are the best technical, ethical, and policy approaches to ensure strong privacy protections while maximizing the benefits of open data?

This briefing paper gives an overview of current approaches to striking a balance between data sharing and privacy protection.[5] It acknowledges that taking privacy into account is a vital piece of any strong

[2] Office of the Press Secretary, "Executive Order -- Establishment of the Federal Privacy Council", The White House, February 9, 2016, https://www.whitehouse.gov/the-press-office/2016/02/09/executive-order-establishment-federal-privacy-council (accessed March 11, 2016).

[3] Emily Shaw, "Exploring open data's microdata frontier", The Sunlight Foundation, October 20, 2014, http://sunlightfoundation.com/blog/2014/10/24/exploring-open-datas-microdata-frontier/ (accessed February 23, 2016).

[4] This paper relies heavily on several sources that focus on the specific intersection of open government data and privacy, but it is built on a broader body of work around data, technology and privacy. The work of Latanya Sweeney and Arvand Narayanan represent two good entry points to the larger topic. For more on differential privacy, which is mentioned briefly near the end of this paper, see the work of Cynthia Dwork.

[5] 1) Kieron O'Hara, "Transparency Government, Not Transparent Citizens: A Report on Privacy and Transparency for the Cabinet Office", School of Electronics and Computer Science, University of Southhampton, 2011 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/61279/transparency-and-

open data initiative. It covers why microdata is important, key issues in providing access to microdata, effective approaches to privacy protection, and further questions to explore.

Why is microdata important?

Analyses of government-held microdata can advance public policy toward societal benefits including greater insight into public issues, better informed decision-making, and better delivery of public services. Microdata is already being used to improve the health and safety of America's citizens, the national transportation infrastructure, the criminal justice system, the quality of education, and the equity and stability of the country's housing market, among other uses. Several current examples show the benefits of releasing microdata under the right conditions - and also reveal the challenges to allowing access to data that contains or could lead to the discovery of sensitive personal information.

Healthcare

A revolution in healthcare is underway, with data at its core. However, given that health data is among the most personal and sensitive of all data, advances in this arena are also demonstrating the challenges of greater data utilization and associated risk assessments. With proper privacy and security mechanisms in place, certain health and medical research institutions are able to share de-identified patient health information with doctors, allowing them to diagnose and treat disease more effectively. Large health datasets have been used to target services to underserved populations.[6] Research centers, drug companies, hospitals, and other institutions can analyze patient data to improve services and develop new treatments.

The Precision Medicine Initiative (PMI) exemplifies the opportunities in analyzing health microdata. Launched in 2015, the PMI is a federal effort to "enable a new era of medicine through research, technology, and policies that empower patients, researchers, and providers to work together toward development of individualized treatments."[7] If successful, it will allow for highly individualized and

---

privacy-review-annex-a.pdf (accessed February 23, 2016). 2) Frederik J. Zuiderveen Borgesius, Mirielle Van Eechoud, Jonathan Gray, "Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework", November 24, 2015, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2695005## (accessed February 23, 2015). 3) Emily Shaw, "How governments are safely opening up microdata", The Sunlight Foundation, November 2, 2015, http://sunlightfoundation.com/blog/2015/11/02/how-governments-are-safely-opening-up-microdata/ (accessed February 23, 2016). 4) John Podesta, Penny Pritzker, Ernest Moniz, John Holdren, Jeffrey Zients, "Big Data: Seizing Opportunities, Preserving Values", Executive Office of the President, May 1, 2014, https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf (accessed February 23, 2016).

[6] Federal Trade Commission, "Big Data: A tool for inclusion or exclusion?", Federal Trade Commission, January 2016, https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf (accessed February 23, 2016), 32.

[7] The White House, "The Precision Medicine Initiatave", The White House, https://www.whitehouse.gov/precision-medicine (accessed February 23, 2016).

targeted treatments based on a range of inputs including personal medical histories and genetic analysis.

The PMI relies heavily on data sharing. As the White House explains, "to get there, we need to incorporate many different types of data... [including] data about the patient collected by health care providers and the patients themselves. Success will require that health data is portable, that it can be easily shared between providers, researchers, and most importantly, patients and research participants."[8]

One of the first priorities of the Precision Medicine Initiative was a set of Privacy and Trust Principles that "...articulate a set of core values and responsible strategies for sustaining public trust and maximizing the benefits of precision medicine." They aim to ensure transparency, strong governance, and data quality while empowering patients and protecting privacy.[9] The principles for data sharing, access, and use, for example, include using methods to preserve the privacy of patients' records, prohibiting unauthorized re-identification of patients, and establishing multiple tiers of data access, from open to controlled, depending on the nature of the data. Overall, the Privacy and Trust Principles outline a strong framework for applying many current approaches to balancing data sharing with privacy.

Transportation

Around the world, untold numbers of commuters now check their mobile phones every day to see when the next bus will arrive. They can thank open data.[10] Ubiquitous travel apps have shown how open transportation data can improve public transit access, ease traffic congestion, and make citizens' lives easier.

On a policy level, transportation microdata can be particularly useful for transit planners. For example, microdata helps planners better understand the obstacles faced by low income workers as they travel to their jobs, allowing for more efficient service delivery and equitable planning decisions.[11]

[8] Ibid.
[9] The White House, "Precision Medicine Initiative: Privacy and Trust Principles", The White House, November 9, 2015, https://www.whitehouse.gov/sites/default/files/microsites/finalpmiprivacyandtrustprinciples.pdf (accessed February 23, 2016), 1.
[10] Elizabeth Press, "A Case for Open Data in Transit", Streetfilms, July 29, 2010, http://www.streetfilms.org/a-case-for-open-data-in-transit/ (accessed February 23, 2016).
[11] Kevin F. Tierney, "Use of the U.S. Census Bureau's Public Use Microdata Sample (PUMS) by State Departments of Transportation and Metropolitan Planning Organizations", Transportation Research Board of the National Acadamies, 2012, http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_syn_434.pdf (accessed February 23, 2016), 30.

Increasingly popular "bike sharing" systems are another example of using transit microdata. These programs generate mountains of data which are often released publicly, allowing advocates to push for expanded service, authorities to better target infrastructure investment, and researchers to ask tough questions about system equality. For example, a recent analysis of 22 million trips taken using New York City's Citi Bike system revealed that the bikes were heavily used for commuting purposes and rides were often concentrated in areas with robust bike lane infrastructure. [12]

Transportation microdata has potentially powerful applications when combined with other types of microdata. At a 2015 roundtable held with the US Department of Transportation and users of its data, participants flagged the need for crash data to be combined with hospital data "to understand the long-term impacts of vehicle crashes and how different kinds of safety equipment can mitigate injury."[13] There are privacy-protection challenges to this type of analysis, but it has the potential to make a significant contribution to public health.[14]

There are also privacy risks in the public release of transit related microdata. To demonstrate this, in 2014 a researcher used data from London's bike sharing system to build a map of trips taken by a single user, revealing very personal information in the process. The researcher did not attempt to identify the specific individual, but posited that it would be easy to do with a small amount of additional information.[15] Another prime example occurred that same year when The New York City Taxi and Limousine Commission released 173 million records of taxi trips from the previous year using cryptographic hashing to de-identify medallion numbers. Researchers were able to crack the code and identify the drivers.[16]

Criminal Justice

Microdata can help improve the criminal justice process at several stages. It can be used to develop effective public policies, improve community relations, and correct unfair practices.

Recent high profile efforts have been focused around opening data about police practices and operations. The Sunlight Foundation has found that previous data releases "have already paid off by

[12] Kelsey E. Thomas, "What 22 Million Rides Tell Us About NYC Bike-Share", Next City, January 28, 2016, https://nextcity.org/daily/entry/citi-bike-new-york-city-bike-share-data (accessed February 23, 2016).
[13] Center for Open Data Enterprise, "Improving Safety Data: A Roundtable with the U.S. Department of Transportation", The Center for Open Data Enterprise, 2015, https://s3.amazonaws.com/odenterprise/DoT+Roundtable+Report.pdf (accessed February 23, 2016), 2.
[14] Ibid, 7-8.
[15] Leo Mirani, "London Transport's Bike-Share Privacy Slip Raises Concerns", Quartz, April 16, 2014, http://skift.com/2014/04/16/london-transports-bike-share-privacy-slip-raises-concerns/ (accessed February 23, 2016).
[16] "Riding with the Stars: Passenger Protection in the NYC Taxicab Dataset," Anthony Tockar, September 15, 2014, http://research.neustar.biz/author/atockar

improving outcomes that communities perceived as unfair. The case of released stop-and-frisk data provides an important example of this, where New York's public release of granular pedestrian stop data, and the analysis it permitted, led to the discovery that almost nine out every 10 people stopped were entirely innocent, and that nine out of every 10 people stopped were non-white."[17] Allowing for better understanding of this data helped kick-start the repeal of an ineffective and discriminatory process.

Releasing microdata about the criminal justice system carries significant privacy concerns. For example, an individual who has been arrested, but never charged with or convicted of a crime, may still have his or her reputation tarnished by the public release of arrest information.[18]

Housing

After the global financial crisis, Congress took a number of steps to safeguard our financial system. Congress mandated the public release of data showing trends in the mortgage industry, in the interest of avoiding another housing bubble. As part of that effort, Congress strengthened requirements for disclosing data under the Home Mortgage Disclosure Act (HMDA), a 1975 law designed to help prevent housing discrimination.

Data collected under HMDA, which is now implemented by the Consumer Financial Protection Bureau (CFPB), is released publicly every September. The data "help show whether lenders are serving the housing needs of their communities; they give public officials information that helps them make decisions and policies; and they shed light on lending patterns that could be discriminatory."[19] Releasing this data comes with privacy risks. To that end, the CFPB has proposed a balancing test "to determine whether and how HMDA data should be modified prior to its disclosure to the public in order to protect applicant and borrower privacy while also fulfilling the disclosure purposes of the statute."[20] Balancing frameworks like this, together with other approaches discussed later in this paper, can be an important part of the solution to releasing microdata while protecting privacy.

---

[17] The Sunlight Foundation, "The Benefits of Criminal Justice Data: Policing and Beyond", The Sunlight Foundation, May 2014, http://assets.sunlightfoundation.com/criminaljustice/sunlight-policy-brief-the-benefits-of-criminal-justice-data-policing-and-beyond.pdf (accessed February 23, 2016), 7.

[18] Damian Ortellado, "Reconciling criminal history open data and expungement", The Sunlight Foundation, February 3, 2016, http://sunlightfoundation.com/blog/2016/02/03/reconciling-criminal-history-open-data-and-expungement/ (accessed February 23, 2016).

[19] Consumer Financial Protection Bureau, "The Home Mortgage Disclosure Act", The Consumer Financial Protection Bureau, http://www.consumerfinance.gov/hmda/learn-more (accessed February 23, 2016).

[20] Consumer Financial Protection Bureau, "Final Rule Home Mortgage Disclosure Regulation, Docket No. CFPB-20140-0019", The Consumer Financial Protection Bureau, 2014, http://files.consumerfinance.gov/f/201510_cfpb_final-rule_home-mortgage-disclosure_regulation-c.pdf (accessed February 23, 2016), 11.

Education

Microdata on student performance can help educational institutions provide students with the tools and support they need to build useful knowledge and skills.[21] To this end, the Obama administration has taken a number of steps to ensure that education data is properly leveraged, and has pledged to "...work to develop a common trust mechanism for schools that want to exchange student data with each other and other qualified parties..."[22] Establishing that trust, however, has been an obstacle to working with student data.

Experience with educational data shows it is critically important for the public to feel that privacy has been considered in the decision making process around data release and sharing.[23] InBloom was a private data analytics company working with educational data from a number of states. The company's goal was to help teachers tailor assignments to better suit the needs of individual students. While "there weren't any documented cases of InBloom misusing the information" that the company held, InBloom did not demonstrate that the company was taking privacy seriously. There was serious pushback from parents and privacy advocates and the company was eventually forced to shut down.[24] This lesson is applicable to government agencies and companies working with sensitive information: They need to take privacy into account early and often, and in a public way, especially moving to "open up" access to data containing PII.

What are the Key Issues to Consider?

In addition to these specific examples, data experts have identified a number of general concerns that apply to most cases where microdata is released.

The Mosaic Effect and Anonymization

The Mosaic Effect is a common term for the idea that disparate datasets and information can be combined to expose sensitive information and negate attempts to protect privacy.[25] There is concern that releasing microdata could result in privacy violations, even if efforts have been made to "anonymize" or "de-identify" it by stripping it of PII. There is also general consensus that there is no

[21] Podesta et al., 63.
[22] Todd Park and Jim Shelton, "The Power of Open Education Data", The White House, June 8, 2012, https://www.whitehouse.gov/blog/2012/06/08/power-open-education-data-0 (accessed February 23, 2016).
[23] O'Hara, 3
[24] Olga Kharif, "Privacy Fears Over Student Data Tracking Lead to InBloom's Shutdown", Bloomberg Business, May 1, 2014, http://www.bloomberg.com/bw/articles/2014-05-01/inbloom-shuts-down-amid-privacy-fears-over-student-data-tracking (accessed February 23, 2016).
[25] Alex Howard, "Open government experts raise concerns about 'mosaic effect' in open data policy", E Pluribus Unum, May 20, 2013, http://e-pluribusunum.org/2013/05/20/open-data-mosaic-effect/ (accessed February 23, 2016).

foolproof way to completely anonymize a dataset, because linking to other sources of data can often give enough information to identify individuals.[26]

Some high-profile examples have fueled these concerns. In what is perhaps the most famous case, researchers were able to identify individuals from supposedly anonymized Netflix rating information a high percentage of the time with only the help of publicly available information from another source, the Internet Movie Database.[27] Another commonly cited example emerged when America Online (AOL) released "anonymized" search results from 650,000 of its users. This turned out to be a case of very weak anonymization, since AOL failed to consider the fact that individuals often perform web searches for their own names, allowing interested individuals to significantly narrow the list of potential names right off the bat.[28]

While the Netflix and AOL examples took place several years ago, they exemplify a continuing concern. There is no broad consensus on the potential risk of the mosaic effect or on the potential and limits of de-identification technology.[29] A later section of this paper, which describes possible solutions, will touch on the current state and applicability of anonymization techniques in the release of microdata.

Existing Law

A number of existing laws relate to privacy and data release. None of them should preclude the government from releasing microdata entirely, but they should be considered to ensure the legality of any open data program. Some key privacy laws include

- The Privacy Act of 1974: Governs how federal agencies collect, maintain, use, and share information about individuals using a set of fair information practices. Requires agencies to give public notice about their systems of records containing PII.[30]
- Health Insurance Portability and Accountability Act: Establishes national standards to protect the privacy of individual medical records and other personal health information.[31]

[26] O'Hara, 47.
[27] Arvind Narayanan and Vitaly Shmatikov "Robust De-anonymization of Large Datasets (How to Break the Anonymity of the Netflix Prize Dataset)", University of Texas at Austin, February 5, 2008, http://arxiv.org/PS_cache/cs/pdf/0610/0610105v2.pdf (accessed February 23, 2015).
[28] Michael Arrington, "AOL Proudly Releases Massive Amounts of Private Data", TechCrunch, August 6, 2006). http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/ (accessed February 23, 2016).
[29] Emily Shaw and Daniel Cloud, "Anonymization and microdata: Can we open up granular info without invading privacy?", The Sunlight Foundation, October 28, 2014, http://sunlightfoundation.com/blog/2014/10/28/anonymization-and-microdata-can-we-open-up-granular-info-without-invading-privacy/ (accessed February 23, 2016).
[30] The Department of Justice, "Privacy Act of 1974", The Department of Justice, http://www.justice.gov/opcl/privacy-act-1974 (accessed February 23, 2016).

- Family Educational Rights and Privacy Act, and Protection of Pupil Rights Amendment: Apply to all schools that receive funds from applicable programs of the US Department of Education. Gives certain rights to parents or children over the age of 18 and protects the privacy of educational records.[32]

Algorithmic Discrimination

The potential for "algorithmic discrimination" has emerged with the growing application of open government data, particularly large datasets containing data on individuals. While this is not a privacy issue per se, it's a concern that can also diminish public trust in government open data programs. Algorithmic discrimination is, essentially, the risk that algorithms for data analysis can inadvertently produce biased results that are unfair to certain groups. Failing to account for this risk during the data release process can "lead decision-makers to discriminate against people who are already more likely to face discrimination, even while these data-based judgements stem less obviously from human prejudice."[33]

In a January 2016 report the Federal Trade Commission (FTC) looked at the issue through the lens of big data.[34] The FTC found companies that rated customers as higher risk because they paid for certain services, such as marriage counseling, with their credit cards. In other cases, customers who paid their credit card bills on time still had their credit limits lowered because they shopped at stores that were frequented by consumers with spotty payment histories.[35] The report noted a number of existing laws designed to prevent discrimination in the private sector, and recommended that similar principles should apply to the use of big data and open data.[36]

At the same time, the FTC report noted that a number of stakeholders believe that "we should recognize the potential benefits of big data to reduce discriminatory harm."[37] If properly applied, microdata --

---

[31] The Department of Health and Human Services, "The HIPAA Privacy Rule", The Department of Health and Human Services, http://www.hhs.gov/hipaa/for-professionals/privacy/ (accessed February 23, 2016).
[32] U.S. Department of Education, "Family Educational Rights and Privacy Act", U.S. Department of Education, http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html (accessed February 23, 2016).
[33] Kate Crawford, The Hidden Biases of Big Data; Harvard Business Review, April 1, 2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data; Danielle Citron and Frank Pasquale. "The Scored Society: Due Process for Automated Predictions." WASHINGTON LAW REVIEW, 2014; Frank PasqualeThe Black Box Society: The Secret Algorithms That Control Money and Information.
[34] Ramirez, Edith, Julie Brill, Maureen Ohlhausen, and Terrell McSweeny. Big Data: A Tool for Exclusion. Report. Washington: FTC, 2016.
[35] Federal Trade Commission, 9.
[36] These include the Fair Credit Reporting Act, the Federal Trade Commission Act, the Equal Credit Opportunity Act, the Americans With Disabilities Act, the Age Discrimination in Employment Act, the Fair Housing Act, the Genetic Information Nondiscrimination Act, and the Equal Right Act of 1964.
[37] Federal Trade Commission, 26.

whether openly released by the government or collected and distributed by private organizations -- can be leveraged to support traditionally underserved populations.

Loss of Public Trust

Public trust and confidence are vital to the success of any data release program. This point has been addressed in several high-profile reports. The 2014 White House report on big data and privacy, released right after InBloom announced that it was shutting down, used educational data as an example. "As students begin to share information with educational institutions," the report said, "they expect that they are doing so in order to develop knowledge and skills, not to have their data used to build extensive profiles about their strengths and weaknesses that could be used to their disadvantage in later years."[38]

Individuals that share information with the government have a reasonable expectation of privacy. In that light, privacy considerations are essential to maintaining public trust in any public release of individual data. As one paper put it, "not only are privacy and transparency compatible, privacy is a necessary condition for a successful transparency programme."[39]

What has worked?

There have been several comprehensive studies of privacy issues related to releasing microdata and approaches for ensuring that open data is both safe and useful. These analyses all agree that there is no one right approach and that policymakers should consider a range of options to protect privacy throughout the open data lifecycle. For that reason, a consistent recommendation is that any open data program should have a strong governance structure.[40] Here are some of the key approaches now being used to balance privacy protection and data release.

De-identification Approaches

The President's Council of Advisors on Science and Technology recently concluded that "anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods."[41] There are mathematical techniques that make it possible to de-identify data

[38] Podesta et al., 63.
[39] O'Hara, 3
[40] Emily Shaw, "How governments are safely opening up microdata", The Sunlight Foundation, November 2, 2015, http://sunlightfoundation.com/blog/2015/11/02/how-governments-are-safely-opening-up-microdata/ (accessed February 23, 2016).
[41] The President's Council of Advisors on Science and Technology, "Big Data and Privacy: A Technological Perspective", Executive Office of the President, May 2014, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf (accessed February 23, 2016), 39.

completely effectively, but at the loss of some accuracy in data analysis. [42] That may be an acceptable tradeoff for some kinds of research but not for others.

Despite their limitations, anonymization techniques can play a role in privacy protection. One report on privacy and transparency in the UK government argued that "It is important to recognise that it would be a terrible defeat if it became impossible to publish useful datasets because of deanonymisation [i.e., identifying data that had been anonymized]."[43] Anonymizing datasets makes it more difficult for potential bad actors to misuse data and may help reduce the privacy risk enough to justify data release in the public interest.

One technique, differential privacy, "aims to maximize the accuracy of database queries or computations while minimizing the identifiability of individuals with records in the database, typically via obfuscation of query results (for example, by the addition of spurious information or "noise")."[44] The Census Bureau has used this approach to release demographic data in its OnTheMap tool.[45]

New applications of cryptography may also provide some privacy protection.[46] However, these are not foolproof, as the New York City Taxi and Limousine Commission discovered in the 2014 example detailed earlier.

De-identification techniques are evolving rapidly, and a full discussion of them is beyond the scope of this paper. A recent paper from the National Institute of Standards and Technology provides a thorough review of types and limits of de-identification.[47]

"Semi-open" Data

It is easier to protect privacy if data is not fully opened to the public, but opened only enough to enable important, selective uses. The Open Data Institute has proposed a Data Spectrum that begins to describe different degrees of openness for different kinds of data.[48]

Government agencies already limit the extent to which certain potentially sensitive data is shared: They may use Memoranda of Understanding to give trusted researchers access to sensitive data on condition that they not share it. As an alternative, agencies can hold data in a "data enclave": They can respond to queries by doing data analysis and providing the results to researchers without actually releasing the

[42] Goroff, Daniel L. "Balancing privacy versus accuracy in research protocols." *Science* 30 Jan 2015.
[43] O'Hara, 48
[44] Ibid, 37.
[45] http://onthemap.ces.census.gov
[46] Ibid, 35.
[47] Garfinkel, Simson. "De-Identification of Personal Information." NISTIR 8053, October 2015. http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf
[48] https://theodi.org/data-spectrum

data. This approach provides greater protection, although it can create an administrative burden for the agency.

Along these lines, Amazon Web Services has worked with government health experts to host a wide array of genetic data in the cloud, allowing researchers and others to use it in a tightly controlled, secure environment.[49] This makes the data more accessible and useful while limiting potential threats associated with its uncontrolled release.

It is also possible to widely release data, but with a "license [that] require[s] re-users to not re-identify data."[50] While these methods aren't traditionally thought of as providing "open data," they can be a good alternative when data is considered too sensitive to share with a completely open license or there are barriers to its wide release.

Legal and Governance Structures

Several laws, regulations, and internationally recognized guidelines provide frameworks for privacy and data protection. It should be noted that none of these were created with open data in mind, but will certainly be helpful when building privacy protections into open data programs.

- **Freedom of Information Laws:** "provide inspiration on how to strike a balance between privacy and transparency in the open data context…[they] typically aim to accommodate privacy interests, for example by reserving access to personal information to parties with particular interests, or by only making records available in secure reading rooms."[51]
- **OECD Privacy Framework**: First published in 1980, the OECD Privacy Guidelines were the first set of internationally agreed upon privacy principles.[52] They were updated and expanded in 2013. The Framework is widely utilized, but has been criticized for their "risk-based approach…[as well as] for promoting business over privacy."[53]
- **Privacy Impact Assessments**: In the United States, government datasets that contain PII must have a corresponding Privacy Impact Assessment. These documents can be useful when balancing the relative costs and benefits of releasing a dataset.
- **Fair Information Practice Principles**: "a set of principles and practices that describe how an information-based society may approach information handling, storage, management, and flows with a view toward maintaining fairness, privacy, and security in a rapidly evolving global

[49] Amazon Web Services, "Genomics in the Cloud", Amazon Web Services, https://aws.amazon.com/health/genomics/ (accessed February 23, 2016).
[50] Borgesius, Eechoud, Gray, 41
[51] Borgesius, Eechoud, Gray, 17-20
[52] Monica Kuschewsky, "Revised OECD Privacy Guidelines Strengthen Accountability Principle", Inside Privacy, September 23, 2013, https://www.insideprivacy.com/international/revised-oecd-privacy-guidelines-strengthen-accountability-principle/ (accessed February 23, 2016).
[53] Borgesius, Eechoud, Gray, 27

technology environment."[54] The Fair Information Practice Principles have been lauded for their "balance [between] privacy-related interests and other interests, such as those of business and the public sector."[55]

In addition to these and other existing laws, privacy experts have suggested several kinds of approaches based on data governance.

Trust frameworks and threat modeling

Data governance approaches make a distinction between "good actors" and "bad actors." When data is released to good actors, such as qualified researchers, re-identification risk can be limited through agreements on conditions of data use. These kinds of agreements can provide a "trust framework" to govern the use of data effectively.

At the same time, trust frameworks are useless against "bad actors" who want to breach privacy protections on purpose. Agencies may want to use "threat modeling" to identify worst-case scenarios and decide what measures they need to prevent them.

Balancing frameworks

Balancing frameworks can provide a method for assessing the risk of broaching privacy against the public good of releasing data. Before *any* dataset is released, the risks to individual privacy should be systematically balanced against the potential value of public disclosure. Balancing frameworks may benefit from a "circumstance catalog" -- a list of " factors that should be considered when assessing whether, and under which conditions, a dataset should be released, as well as different options for how it should be released" -- to help guide decision making about data release under certain conditions.[56]

Many Freedom of Information laws strike a balance between privacy and openness, but there are fewer laws or regulations that apply this approach to proactive open data programs.[57] In the United States, the CFPB's proposed balancing test for the release of HMDA data (described above) is one of the first examples of such an approach in a federal rulemaking, and may become a model for others.[58] Additionally, balancing frameworks can benefit from demand-driven processes where requests for specific datasets must be justified in a way that balances the public value of the data with the need for privacy protection.[59]

[54] https://www.worldprivacyforum.org/2008/01/report-a-brief-introduction-to-fair-information-practices/
[55] Borgesius, Eechoud, Gray, 23
[56] Borgesius, Eechoud, Gray, 42.
[57] Ibid, 20.
[58] Consumer Financial Protection Bureau, 11.
[59] O'Hara, 62.

"Metatransparency"

It's become an accepted best practice for open data programs to build transparency into their governance structures.[60] The United States government releases robust metadata about agency datasets, including relevant information about the potential privacy impacts of data release[61]. Different thresholds of privacy may be acceptable for datasets with different kinds of personal information. The ability to learn which college someone attended, for example, may be less of a risk than accessing his or her medical records. Metadata can help identify the source, nature, and privacy risk posed by different kinds of data.

Ongoing Evaluation

Privacy protection is an evolving field. Today's approaches may no longer be best practices in five years as new ways to protect privacy are developed. The Open Government Guide's chapter on "privacy and data protection" lays out steps that governments can take at the initial, intermediate, advanced, and innovative stages of open data programs.[62] Governance structures can be developed to be flexible enough to respond to new threats and incorporate new research and methodology.[63]

Questions for Further Consideration

This briefing paper sets the stage for a broader conversation around open data and privacy. Some key considerations for further discussion include:

- What is known today about the pros and cons of the approaches described in this paper? What more needs to be learned in order to apply them broadly?
- Who and what are the best sources for information, expertise, and implementation?
- How can we best learn from organizations with experience and expertise in these areas? Can we generalize from approaches used by one agency – or by a company or nonprofit – to develop scalable solutions across government?
- How can we address privacy issues most efficiently across government? Are there ways to lower costs or share them among agencies?

Acknowledgments

The Center for Open Data Enterprise thanks Open Data Partner Microsoft and Open Data Supporter Booz Allen Hamilton for supporting the Center's work on the Open Data Roundtables and this Briefing

[60] Ibid, 35.
[61] Labs.Data.Gov, "Project Open Data Dashboard", Data.gov, http://labs.data.gov/dashboard/offices, (accessed February 23, 2016).
[62] Open Government Guide, "Privacy and data protection", The Open Government Guide, http://www.opengovguide.com/topics/privacy-and-data-protection/ (accessed February 23, 2016).
[63] O'Hara, 69-76

Paper. The lead researcher and author for this paper is Matthew Rumsey, Research Fellow at the Center for Open Data Enterprise. The Center is an independent nonprofit 501(c)3 organization, based in Washington, DC, whose mission is to maximize the value of open data as a public resource.

Copyright Statement

Further Reading

This briefing paper aims to present various examples and analysis that may be relevant to a discussion around the privacy implications of open data; it is not intended as a comprehensive analysis of the topic. The following sources provide a deeper look at the issues discussed here.

Executive Order – Establishing the Federal Privacy Council

https://www.whitehouse.gov/the-press-office/2016/02/09/executive-order-establishment-federal-privacy-council

The Sunlight Foundation's series on opening up microdata and Criminal Justice data

http://sunlightfoundation.com/blog/tag/opendata1/

http://sunlightfoundation.com/criminaljustice/

Big Data: Seizing Opportunities, Preserving Values - White House report on big data and privacy

https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

Big Data and Privacy: A Technological Perspective - Accompanying report by the President's Council of Advisors on Science and Technology

https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

Big Data: A tool for inclusion or exclusion? - Federal Trade Commission report

https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf

Exploring data-driven innovation as a new source of growth: Mapping the policy issues raised by "big data" – OECD

http://www.keepeek.com/Digital-Asset-Management/oecd/industry-and-services/supporting-investment-in-knowledge-capital-growth-and-innovation/exploring-data-driven-innovation-as-a-new-source-of-growth-mapping-the-policy-issues-raised-by-big-data_9789264193307-12-en#page1

De-identification of Personal Information – NIST Report

http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

OPEN DATA, PRIVACY, AND FAIR INFORMATION PRINCIPLES: TOWARDS A BALANCING FRAMEWORK

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2695005

The Open Government Guide on Privacy and Data Protection

http://www.opengovguide.com/topics/privacy-and-data-protection/