October 2, 2017

Dear Kelsey Finch,

We are pleased to have the opportunity to provide comments on the Future of Privacy Forum proposed draft report for the City of Seattle Open Data Risk Assessment. Municipalities face significant challenges regarding the design and implementation of privacy-aware open data systems. A model risk assessment framework to guide decisions in this area is greatly needed.

The proposed draft report outlines a number of recommendations that are likely to help guide cities as they carry out risk assessments for their open data programs. In particular, we support the following recommendations and observations in the draft report that reflect concepts from a modern approach to privacy protection:

- The emphasis on a lifecycle approach to data management, and, in particular, the recognition of the need for controls at the collection stage (e.g., "Because of the interplay of open data and public records requests, municipalities must be far-sighted in deciding what data they will collect in the first place.") (p. 8),
- References to a range of available technical, legal, and procedural controls for privacy protection (Appendix B),
- The City of Seattle's open data policy's requirement that privacy risk assessments be conducted on an annual basis (p. 5),
- The observation that "it is no longer always clear when data is 'personally identifiable'" and that "data that was once non-identifiable may become identifiable over time" (p. 7),
- The recommendation that "municipalities should not rely on static lists of PII in determining if a particular dataset creates a risk of re-identification" (p. 7),
- The observation that "once information has been published publicly, it likely can never be retracted" (p. 8), and
- The observation that "if the data exposes vulnerable populations to higher privacy risks or at a higher rate than others, it may be inequitable" (p. 10).

Our comments focus on the following opportunities for strengthening the report:

- The proposed risk assessment framework is expressly intended to be applied by "open data managers and departmental data owners" and "without a bevy of expert statisticians,

privacy lawyers, or philosophers" (p. 3). Many of the technical measures described in this report, including suppression, generalization, pseudonymization, aggregation, and k-anonymity can sometimes provide reliable privacy protection, if applied by expert statisticians. Statisticians at federal statistical agencies who specialize in disclosure avoidance, for example, are equipped to apply sophisticated disclosure limitation techniques to mitigate privacy risks before releasing data to the public. However, the techniques they use cannot readily be applied effectively by non-experts. It is not reasonable to expect open data managers without statistical training or expertise in privacy to engage in ad hoc, "flexible, risk-based" decisions that will effectively "maximize the utility and openness of civic data while minimizing privacy risks to individuals and community concerns about ethical challenges, fairness, and equity," as the draft report promises.

- The real-world examples highlighted in the report, including the release of the names of sexual assault victims by the City of Dallas open data portal and the release of sensitive information about gun permit applicants by the City of Philadelphia, demonstrate how a city can adopt a reasonable open data policy that requires reviewing, classifying, and suppressing sensitive information, yet still fall short of robust privacy protection when applied in an ad hoc fashion. The report should clarify how the model risk assessment framework should guide the design of a systematic open data management program that aims to prevent case-by-case determinations by open data managers that can inadvertently lead to similar types of disclosures. In particular, approaches such as differential privacy, which do not rely on data managers to specify particular fields as sensitive, could be recommended as part of a comprehensive open data management plan. More broadly, the recommendations should provide concrete guidance on how to apply the risk assessment framework so as to safeguard against the types of disclosures seen in the Dallas and Philadelphia examples.
- The examples from Dallas and Philadelphia also illustrate how unstructured data (such as free-form text fields) carry heightened privacy risks. The state-of-the-art for transforming unstructured data is not sufficient to prevent re-identification or leakage of sensitive information. This report could recommend that, when collecting unstructured data, individuals should be provided with clear notice that the information they provide will be made available to the public. Individuals should be provided options, such as the ability to opt out of data collection, to designate their information as sensitive and to be withheld from future release, and to review and correct their records in the future. In cases in which individuals are not able to opt out of data collection, or in which opting out would be too costly for some individuals, the information collected should be subject to stronger controls.

- Because robust privacy management requires highly technical expertise, the report should also provide details regarding when and how to seek guidance from technical experts on designing privacy-aware data releases. Experts can be consulted through a process similar to an institutional review board or privacy board for two purposes: for the design of data release programs on a one-time or periodic basis, and on a case-by-case basis for specific data release decisions. Because there is a small (but growing) pool of experts to draw from, this review board could be created as a centralized resource available for consultation by open data managers across the country. The role of experts is to evaluate tradeoffs and design efficient mechanisms for the best tradeoff between privacy and information sharing. The choice of tradeoffs is a political and social one that requires assessing the value of the information for transparency and accountability. This choice of tradeoffs should be made in the open, through a participatory process that involves multiple stakeholders, including privacy and open government advocates.

- The report could encourage the adoption of formal privacy models such as differential privacy which ensure consistent and robust privacy protection at a large scale without relying on ad hoc case-by-case determinations. Because the cost associated with each city developing and implementing its own differentially private tools may be prohibitive, the report could recommend that cities request that vendors supplying open data portal software include differentially private tools in their platforms and work with them to define the design requirements for such tools. In the absence of determinations based on specialized data privacy expertise and general tools for privacy protection such as tools that satisfy differential privacy, cities should be encouraged to strongly err on the side of caution when releasing data to the public.

- The draft report recommends that cities address harms related to fairness and observes that "'residents of zip codes listed as having high rates of households below the poverty level; property owners in neighborhoods where crime rates are higher than average; [and] students at schools that are underperforming' may all be adversely effected [sic] by conclusions drawn from such datasets" (p. 10). Fairness is a challenging problem, and, although the academic literature proposes a number of measures of fairness, a consensus on how to overcome the limitations and tradeoffs between these measures has not yet emerged. Tasking open data managers with making ad hoc decisions regarding fairness concerns when releasing population-level statistics, without consulting with experts in this domain, is a tall order. Therefore, to address concerns related to fairness, a review board of ethicists and technical experts with experience in information sharing could be assembled and engaged for case-by-case issue spotting. The choice of solution is a political and social one, requiring openness to the public and participation by multiple stakeholders.

- The draft report observes that "if the data exposes vulnerable populations to higher privacy risks or at a higher rate than others, it may be inequitable" (p. 10). The report could recommend approaches such as differential privacy to open data managers who seek to address "worst-case" privacy risks. Differential privacy offers advantages over heuristic approaches to privacy, such as suppression, generalization, and aggregation which, at best, provide lower bounds on privacy risk and potentially expose certain members of the population, especially minority populations, to higher risks than the general population.

- The draft report emphasizes re-identification risks, but it should also address other privacy risks beyond re-identification enabled by record linkage to known sources of auxiliary information. Technological advances are enabling new and sophisticated attacks, such as statistical inference attacks, that were unforeseen at the time that many current approaches and standards were adopted. Computer scientists recognize the need to protect against not only known modes of attack, but also unknown future attacks. Emerging mathematical approaches, such as differential privacy, can help provide strong privacy protection guarantees that are effective against a very wide range of attacks on privacy, including currently unforeseen ones.

- The report acknowledges that "data de-identification is a moving target," as "data that could not be linked to an individual when it was released could become identifiable over time" (p. 8). This observation supports the use of technologies that satisfy as differential privacy, as this is the only known approach that can address unforeseen classes of attacks, attacks leveraging unanticipated sources of auxiliary information, and attacks enabled by the cumulative privacy risk from multiple analyses based on the same individuals. The report could explicitly acknowledge the benefits of differential privacy to address these risks.

- The discussion of a potential adversary ("whether an expert skilled in re-identifying individuals from seemingly 'anonymous' information, or a commercial information reseller with access to millions of other data points, or an insider who knows other personal information") could be made more concrete by referring to specific examples. For instance, the last category could list family members, friends, coworkers, and neighbors and the kinds of knowledge they may possess about a person described in a municipal open data portal, illustrating how re-identification may in many cases be trivial for such an adversary.

- The report's discussion of expected uses of data should distinguish between population-level and individual-level analyses. The design of a city's data management program should aim to tailor data releases to the risks and intended uses of the data, and make

different datasets available through different mechanisms depending on the relevant categories of intended uses. There is an opportunity to provide detailed guidance to cities as they consider the proper modes of release for different anticipated data uses. For example, individual-level data may be required for expected uses of business license records, but analysts of crime statistics may require only aggregate-level data.

- In describing the information lifecycle, the draft report refers to the stages "from collection to use to share to deletion" (p. 11), but it is important to clarify that the lifecycle does not end at deletion. Rather, the information lifecycle is properly viewed as a cycle that persists through long-term storage and re-use. Although many data management plans rely on data destruction as a technique for protecting privacy, this approach alone should not be considered sufficient for eliminating risk, as deleting data does not mitigate all risks if the data have previously been used or shared.

- In the collection of "high-risk" datasets designated for close review in the proposed risk assessment (p. 16), 311 constituent request records could be added as a category of data at high risk of enabling sensitive information about individuals to be inferred.

- The description of differential privacy provided in Appendix B (p. 34) contains inaccuracies. For instance, differential privacy is not a "set of techniques" but a definition that a wide range of technologies can be designed to satisfy. The differential privacy guarantee does not expire after "a certain period of time." Additionally, differential privacy is not limited to interactive, or query-based, mechanisms as stated in the report. Various techniques, both interactive and non-interactive, can be rigorously shown to satisfy the differential privacy definition. Government agencies such as the Census Bureau and corporations such as Google use differential privacy to provide strong privacy protection when sharing statistics. In particular, the Census Bureau makes data available using a non-interactive differentially private mechanism. Additional tools for differentially private analysis, including tools that are broadly-applicable and can be integrated with a wide range of existing software platforms, are under development at a number of research institutions.

- Appendix B should contain more detailed guidance, particularly regarding how the results of the risk-benefit assessment map to specific privacy controls. The article, Micah Altman et al., Towards a Modern Approach to Privacy-Aware Government Data Releases, 30 Berkeley Tech. L. J. 1968 (2015), https://cyber.harvard.edu/publications/2016/Privacy_Aware_Government_Data_Releases , provides a framework for analyzing the desired data uses and expected benefits, and examining each stage of the data lifecycle to identify specific privacy threats, harms, and vulnerabilities. This report could aim to provide detailed guidance materials to help open

data managers systematically evaluate the privacy risks associated with their activities, as well as select and calibrate privacy and security controls that are suitable for mitigating those risks while enabling the data uses they intend to support.

- In particular, Appendix B could include guidance on implementing a tiered access system. A tiered access model is one in which data are made available to different categories of data users through different mechanisms. For example, a city could provide public access to some data without restriction after robust disclosure limitation techniques have transformed the data into differentially private statistics. Data users who intend to perform analyses that require the full dataset, including direct and indirect identifiers, could be instructed to submit a request, and their use of the data would be restricted by the terms of a data use agreement. Where appropriate, a tiered access model can be used to closely match controls to different risks and intended uses at each stage of the information lifecycle.

More detailed redline comments on the draft report can be found in the enclosed document.

Sincerely,

Alexandra Wood
Berkman Klein Center for Internet & Society, Harvard University

Micah Altman
MIT Libraries, Massachusetts Institute of Technology

Suso Baleato
Institute for Quantitative Social Science, Harvard University

Salil Vadhan
John A. Paulson School of Engineering and Applied Sciences, Harvard University

City of Seattle
Open Data Risk Assessment

JULY 2017 – PROPOSED DRAFT REPORT

DRAFT REPORT: Table of Contents

DRAFT REPORT: Executive Summary

The transparency goals of the open data movement serve important social, economic, and democratic functions in cities like Seattle. At the same time, some municipal datasets about the city and its citizens' activities carry inherent risks to individual privacy when shared publicly. In 2016, the City of Seattle declared in its Open Data Policy that the city's data would be "open by preference," except when doing so may affect individual privacy.[1] To ensure its Open Data program effectively protects individuals, Seattle committed to performing an annual risk assessment and tasked the Future of Privacy Forum (FPF) with creating and deploying an initial privacy risk assessment methodology for open data.

This Draft Report provides tools and guidance to the City of Seattle and other municipalities navigating the complex policy, operational, technical, organizational, and ethical standards that support privacy-protective open data programs. Although there is a growing body of research on open data privacy, open data managers and departmental data owners need to be able to employ a standardized methodology for assessing the privacy risks and benefits of particular datasets internally, without a bevy of expert statisticians, privacy lawyers, or philosophers. By following a flexible, risk-based assessment process, the City of Seattle – and other municipal open data programs – can maximize the utility and openness of civic data while minimizing privacy risks to individuals and community concerns about ethical challenges, fairness, and equity.

This Draft Report first describes inherent privacy risks in an open data landscape, with an emphasis on potential harms related to re-identification, data quality, and fairness. Accompanying this, the Draft Report includes a Model Open Data Benefit Risk Analysis (MODBRA). The model template evaluates the types of data contained in a proposed open dataset, the potential benefits – and concomitant risks – of releasing the dataset publicly, and strategies for effective de-identification and risk mitigation. This holistic assessment guides city officials to determine whether to release the dataset openly, in a limited access environment, or to withhold it from publication (absent countervailing public policy considerations). The Draft Report methodology builds on extensive work done in this field by experts at the National Institute of Standards and Technology, the University of Washington, the Berkman Klein Center for Internet & Society at Harvard University, and others,[2] and adapts existing frameworks to the unique challenges faced by cities as local governments, technological system integrators, and consumer facing service providers.[3]

---

[1] Exec. Order No. 2016-01 (Feb. 4, 2016), *available at* http://murray.seattle.gov/wp-content/uploads/2016/02/2.26-EO.pdf.

[2] *See infra* Appendix A for a full list of resources.

[3] *See* Kelsey Finch & Omer Tene, *The City as a Platform: Enhancing Privacy and Transparency in Smart Communities*, CAMBRIDGE HANDBOOK OF CONSUMER PRIVACY (forthcoming).

Following a period of public comment and input on the Draft Report and proposed methodology, a Final Report will assess the City of Seattle as a model municipality, considering its open data program across six domains:

1. Privacy leadership and management
2. Benefit-risk assessments
3. De-identification tools and strategies
4. Data quality
5. Data equity and fairness
6. Transparency and public engagement

The Final Report will conclude by detailing concrete technical, operational, and organizational recommendations to enable the Seattle Open Data program's approach to identify and address key privacy, ethical and equity risks, in light of the city's current policies and practices.

The City of Seattle is one of the most innovative cities in the country, with an engaged and civic-minded citizenry, active urban leadership, and a technologically sophisticated business community. By continuing to complement its growing open data program with robust privacy protections and policies, the City of Seattle will be able to fulfill its goals, supporting civic innovation while protecting individual privacy in its Open Data program.

Acknowledgments: We extend our thanks to the experts from the City of Seattle, Seattle Community Technical Advisory Board, University of Washington, Berkman Klein Center for Internet & Society at Harvard University, members of the FPF Smart City Privacy Working Group, and others who provided their support and input in the development of this draft report. Special thanks to Jan Whittington, Meg Young, Ryan Calo, Mike Simon, Jesse Woo, and Peter Schmiedeskamp for their foundational scholarship and to Michael Mattmiller, Jim Loter, David Doyle, and the many Open Data Champs for their vision and dedication to making open data privacy a reality for the City of Seattle.

Background

In February 2016, City of Seattle Mayor Edward Murray issued an Executive Order calling for "all city data to be 'open by preference' – meaning city departments will make their data accessible to the public, after screening for privacy and security considerations."[4] The Executive Order "both sets the expectation that public data will be public and makes clear that [the city] has a responsibility to protect privacy."[5]

The City of Seattle Open Data Policy[6] directs the City of Seattle to perform an annual risk assessment of both the Open Data Program and the content available on the Open Data Portal. For this, the City of Seattle contracted the Future of Privacy Forum to develop a methodology for conducting a risk assessment and to actively deploy the methodology. FPF will review a subset of high-risk agency datasets as well as a random sample of additional agency datasets, to evaluate privacy risks, including of re-identification, in case of release of individual datasets or multiple datasets.

From fall 2016 through summer 2017, FPF studied existing privacy and other risk assessment frameworks, created the Model Open Data Benefit Risk Analysis, and assessed the inherent privacy risks in the municipal open data landscape for the City of Seattle as a model municipality. In doing so, FPF built on open frameworks, such as the National Institute of Standards and Technology (NIST) Special Publication 800-series. In addition to a review of available research and policy guidance related to open data privacy risk, FPF conducted interviews with privacy, open data, and disclosure control experts from around the world.

FPF also visited on-site to conduct interviews with Seattle IT and Open Data leadership, departmental Open Data and Privacy Champions, and local community advisors. These interviews included teams from the Seattle IT, Seattle Police Department, Seattle Department of Transportation, Planning and Development, Parks and Recreation, Civil Rights, Immigrant Affairs, and the Seattle Public Library.

FPF presented an early draft of the identified privacy risks and assessment methodology to the Seattle Community Technology Advisory Board (CTAB) for review and input in February 2017. An additional 45-day period for public comment on the report will be offered from July through September 2017.

---

[4] Exec. Order No. 2016-01 (Feb. 4, 2016), *available at* http://murray.seattle.gov/wp-content/uploads/2016/02/2.26-EO.pdf.

[5] CITY OF SEATTLE 2017 OPEN DATA PLAN, http://www.seattle.gov/Documents/Departments/SeattleIT/City%20of%20Seattle%202017%20Open%20Data%20Plan.pdf.

[6] CITY OF SEATTLE, OD-1 V1.0, OPEN DATA POLICY (§ 5(k)) (2016), *available at* http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPolicyV1.pdf.

Open Data Privacy Risks

Open and accessible public data can benefit individuals, companies, communities, and government by unleashing new social, economic, and civic innovations and improving government accountability and transparency. Tremendous benefits in healthcare, education, housing, transportation, criminal justice, and public safety are already being realized as richer and more timely datasets are made available to the public. Open data can unite the power of city and private sector abilities to improve community health and lifestyles, including everything from bikeshare systems and commercial apps harnessing transit data to community advocates shining the light on ineffective or discriminatory practices through policing and criminal justice data.

In Seattle, for example, the Open Data program seeks to:

- "Improve public understanding of City operations and other information concerning their communities,
- Generate economic opportunity for individuals and companies that benefit from the knowledge created by Open Data,
- Empower City employees to be more effective, better coordinated internally, and able to identify opportunities to better serve the public, and
- Encourage the development of innovative technology solutions that improve quality of life."[7]

However, it can also pose substantial risks to the privacy of individuals whose information is collected and shared by the city. Inadequate privacy protections for open data can lead to significant financial, physical, reputational, organizational, and societal harms.

Cities must be vigilant and resourceful to deter and defend against these privacy risks, no matter how they arise. In this section, we describe the core privacy risks facing municipal open data programs: re-identification, biased or inaccurate data, and loss of public trust.

Re-identification

One of the principal and unavoidable risks of opening government datasets to the public is the possibility that the data might reveal private or sensitive information about a specific individual. In cases where open datasets are not adequately vetted, personally identifiable information (PII) may be

[7] *Open Data Program*, CITY OF SEATTLE, https://data.seattle.gov/stories/s/urux-ir64 (last visited July 6, 2017).

published inadvertently. Even when a dataset has been scrubbed of names and other potentially identifying traits and rendered "de-identified," there is a chance that someone (referred to in professional literature as an "adversary")– whether an expert skilled in re-identifying individuals from seemingly "anonymous" information, or a commercial information reseller with access to millions of other data points, or an insider who knows other personal information – might be able to deduce that some of the data relates to a specific individual.

Re-identifying a person in this way not only exposes data about the individual that would otherwise not be available to the public, but could potentially carry embarrassing, damaging, or life-threatening implications. For example, in Dallas, the names of six people who complained of sexual assault were published online by the police department. While the Dallas Police Department does not intentionally publish such sensitive information, of course, its case classification scheme and overlapping information across datasets combined in such a way that the six injured parties could be singled out and identified when they should not have been.[8] Other re-identification attacks may reveal an individual's home address or place of work, exposing them to increased risk of burglary, property crime, or assault.[9]

Recent advances in smart city technologies, re-identification science, data marketplaces, and big data analytics have enhanced re-identification risks, and thus increased the overall privacy risk in open datasets. As open data programs mature and shift from merely providing historic data and statistics to more granular, searchable, accessible, and comprehensive "microdata" about citizens and their activities, the risk of re-identification rises even further. Databases of calls to emergency services, civil complaints about building codes and restaurants, and even civil rights violations will potentially become available for anyone in the world to explore. The ease at which adversaries (including professional researchers, commercial organizations and data brokers, other government and law enforcement agencies, civic hackers, and individual members of the general public) can download, re-sort, and recombine these datasets carries an obvious risk for the leakage of sensitive data.

Even as open data programs take on the challenges of sophisticated re-identification adversaries combining multiple databases to reveal sensitive attributes about individuals, datasets that appear more bureaucratic or even mundane and therefore fail to raise the same privacy red flags – could ultimately leave individuals exposed. In 2017, for example, a parent who was examining expenditure files on the Chicago Public School's website discovered that deep within the tens of thousands of rows of vendor payment data were some 4,500 files that identified students with Individualized Educational Programs – revealing in plain text the students' names, identification numbers, the type of special education

[8] *See* Andrea Peterson, *Why the names of six people who complained of sexual assault were published online by Dallas police,* WASH. POST, Apr. 21 2016, https://www.washingtonpost.com/news/the-switch/wp/2016/04/29/why-the-names-of-six-people-who-complained-of-sexual-assault-were-published-online-by-dallas-police/.
[9] *See* SIMSON L. GARFINKEL, DE-IDENTIFYING PERSONAL INFORMATION NISTIR 8053 (NIST Oct. 2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

services that were being provided for them, how much those services cost, the names of therapists, and how often students met with the specialists.[10]

One of the unavoidable challenges of open data is that once information has been published publicly, it likely can never be retracted. Unfortunately, data de-identification is a moving target – data that could not be linked to an individual when it was released, could become identifiable over time. For example, if sometime in the future another dataset is published that links one record to another or if a new technique becomes available to match information across multiple datasets, the difficulty of re-identifying an individual in the original open dataset may drop significantly. While it is difficult to predict when such future data may become available, cutting-edge research into more dynamic de-identification techniques is underway among disclosure control experts and at statistical agencies around the world.

Re-identification also harms municipalities: when data published on an open data program becomes re-identified and harms an individual, public trust in the city and in open data is seriously eroded. Citizens may stop providing data, or provide false data, if they believe that it might be exposed in the future. If the data were subject to regulatory or confidentiality provisions, moreover, such disclosures could lead to new compliance costs or lawsuits. For example, in 2012, Philadelphia's Department of Licenses & Inspections published gun permit appeals as part of its open data initiative. These permits included a free text field where applicants explained why they needed the permit. Some individuals wrote they carried large sums of cash at night. As a consequence of disclosing this information, the City was ultimately charged $1.4 million as part of a class-action lawsuit. One of the lawyers behind the suit stated that the information released was a "road map for criminals."[11]

Re-identification can cause harms to individuals, to organizations and government agencies, and to society as a whole. Even *false* claims of re-identification can cause significant damage, leaving individuals uncertain whether their information is exposed and susceptible to lost opportunities or mistaken decisions based on data wrongly attributed to them.

Data Quality

Multiple stakeholders rely on the accuracy of information in public datasets: citizens, companies, community organizations, and other governmental entities. In some circumstances, inaccurate, incomplete, or biased open data may have little impact – for example, a list of sold city fleet vehicles may accidentally record the wrong make and model for a vehicle or two. In other circumstances,

[10] *See* Lauren Fitzpatrick, *CPS privacy breach bared confidential student information,* CHI. SUN-TIMES (Feb. 2, 2017), http://chicago.suntimes.com/news/cps-privacy-breach-bared-confidential-student-information/.
[11] *See* Vince Lattanzio, *Philly paying $1.4 million after posting confidential gun permit information online,* NBC PHILADELPHIA, July 22, 2014, http://www.nbcphiladelphia.com/news/local/Philly-Paying-14M-After-Posting-Confidential-Gun-Permit-Information-Online-268147322.html.

however, the consequences can be more lasting, leading to poor or inefficient decision-making, unethical or illegal data uses, or discriminatory outcomes. Publishing the wrong person's information to an open dataset of DUI arrests, for example, could adversely affect that person's employment, credit, and insurance prospects for years to come. Because open data is used so widely and for so many diverse purposes, it is critical that any data released be accurate and unbiased.[12]

Personal data that has been made public without legal conditions may be consumed and repurposed by any number of potential actors, including identity thieves, commercial information resellers (and ultimately their clients, including potential employers, insurers, creditors, and others), companies, friends and family, nosy neighbors, stalkers, law enforcement and other government entities, and others. Some commercial "mugshot" or arrest record databases, for example, profit by gathering sensitive personal information via public records, publishing the data to private sites, and then charging individuals a fee to have them removed.[13] The lack of control over downstream uses of open data is a significant point of concern among a variety of open data stakeholders, including civic hackers, legal advocates, and industry representatives.[14]

Over the last few years, organizations increasingly rely on data to automate their decision-making in a wide variety of situations, including everything from traffic management to personalized advertising to insurance rate setting. But particularly in "smart" systems that use algorithmic decision-making and machine learning, bad data can lead to bad policies. For example, both predictive policing and criminal sentencing have repeatedly demonstrated racial bias in both the inputs (historic arrest and recidivism data) and their outputs, leading to new forms of institutional racial profiling and discrimination.[15]

In fact, even individuals who are not directly represented in an open dataset may nevertheless be impacted by inaccuracies and biases in the dataset or analysis performed on it.[16] For example, according to the City of Seattle, "residents of zip codes listed as having high rates of households below the poverty level; property owners in neighborhoods where crime rates are higher than average; [and] students at schools that are underperforming" may all be adversely effected by conclusions drawn from such

[12] Ironically, the process of de-identifying data to be released publicly may introduce bias or inaccuracies into the dataset. SIMSON L. GARFINKEL, DE-IDENTIFYING GOVERNMENT DATASETS SP 800-188, at 16 (NIST draft. Aug. 2016), http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf.
[13] Damian Ortellado, *The perils of personally identifiable pre-conviction data*, SUNLIGHT FOUNDATION (Feb. 1, 2016, 3:48 PM), https://sunlightfoundation.com/2016/02/01/the-perils-of-personally-identifiable-pre-conviction-data/.
[14] Jan Whittington et al., *Push, Pull, and Spill: A Transdisciplinary Case Study in Municipal Open Government*, 30 BERKELEY TECH. L.J. 1899, 1913-14 (2015).
[15] *See generally* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[16] *See* SIMSON L. GARFINKEL, *supra* note 8.

datasets, especially if drawn from low-quality data.[17] These sorts of inferential disclosures may result in group harms that have not been traditionally viewed as privacy concerns, and may thus not be well addressed by existing municipal privacy policies and practices.

Moreover, an unfair distribution of data benefits and data risks across a community may reinforce societal biases, disguise prejudiced decision-making, and block equal opportunities for marginalized or vulnerable populations. Some open data stakeholders have raised concerns that, particularly when commercialized, public municipal data may be used to "lower property values, redline insurance, et cetera, in neighborhoods with high crime rates rather than addressing those issues."[18] If data represented on the open data program is disproportionately collected from certain populations over others, or is used against certain populations over others, or if the data exposes vulnerable populations to higher privacy risks or at a higher rate than others, it may be inequitable. For example, given that minority and vulnerable populations, including immigrant communities, tend to be over-surveilled in comparison to majority populations, particularly in the context of law enforcement and social services, they may be disproportionately represented in open datasets, creating fertile grounds for inaccuracies and biases in decision making or even just reporting of data. Governments must constantly strive to serve all their citizens fairly and equitably, however difficult it may be to strike the balance of equities.

Public Impact

Open data programs cannot succeed in their social, economic, and democratic missions without public trust. Where individuals feel their privacy is violated by a particular dataset being published or that public expectations of privacy were disregarded, they will hold the open data program accountable. This can result not only in a loss of trust in the open data program, but also in undermining the entire city government's ability to act as a responsible data steward. [19] Civic engagement and communication, paired with demonstrable responsible data practices, can earn the public's trust in open data. But if the public's trust in a government as a responsible data steward is damaged, individuals may become unwilling to support and participate in important civic activities and research.[20] It can also lead to the public providing false data in certain circumstances out of a fear their real information would be compromised.

---

[17] *See* CITY OF SEATTLE, OPEN DATA PLAYBOOK V. 1.0, http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPlaybook_Published_2016.08.pdf.

[18] Whittington et al., *supra* note 13, at 1919.

[19] *See Ben Green et al*., OPEN DATA PRIVACY (2017), https://dash.harvard.edu/handle/1/30340010; Whittington et al., *supra* note 13, at 1914.

[20] SEAN A. MUNSON ET AL., ATTITUDES TOWARD ONLINE AVAILABILITY OF US PUBLIC RECORDS (2011), https://pdfs.semanticscholar.org/fa4b/e73719e5047fb97f21eef25bbe26984abbf0.pdf.

Just as in the event of a data breach, individuals who believe that their personal data may have been exposed to the world will feel uncertainty and anxiety about the loss of informational control and potential long-term ramifications such as identity theft. When personally identifiable information is published to an open data program or a re-identification attack appears successful, individuals often have little recourse. Municipal leaders must be aware that deciding what data they may *release* about individuals is inextricable from what data they *collect* about individuals. Failing to address privacy throughout the entire data lifecycle, from collection to use to sharing to deletion, will impede public trust in data-driven municipal programs. For example, cities should be cautious about collecting information that would harm individuals if it were one day shared via the open data program, disclosed via a public records request, or exposed via a data breach.[21]

Finally, cities must be aware that *how* data is collected and used is as important as how it is released for ensuring public trust in open data programs. Cities must communicate clearly with individuals about how and when their data can find its way to an open data portal. Vague privacy notices and a lack of an opportunity to opt in or out of data collection may shock or surprise some people, even if that information is in pseudonymized or aggregate form. And if data is used for a purpose other than the reason the collection occurred without citizens' consent to repurpose, significant privacy concerns are raised, as well as ethical and technical questions. It is possible that an individual never would have consented to the data collection if they it would ultimately be released through the open data program. Where an individual's privacy – or trust – has been violated by a government data program once, it may be impossible to restore.

*

The transparency goals of municipal open data programs are critical to the improvement of civic life and institutions in the modern city, and rely on the release of microdata about the city and its citizens' activities. And yet people who provide personal information to their governments must be able to trust that their privacy will be protected. If individuals find their personal information exposed, or their neighborhoods singled out or discriminated against, or their data collected for one purpose and used for another, this can undermine public trust in the city as a whole and slow or even reverse the momentum of the open data program. On the other hand, where cities engage the public and communicate the benefits of the open data program while clearly addressing any shortcomings, they may build public trust. Responsible privacy practices and effective communication provide the foundation for successful, trustworthy, and innovative open data programs.

[21] *See* Liz Robbins, *New York City ID Holders Aren't a Threat, N.Y.P.D. Official Says in Court*, N.Y. TIMES (Jan. 5 2017), https://www.nytimes.com/2017/01/05/nyregion/new-york-id-program-immigrants.html?action=click&contentCollection=N.Y.%20%2F%20Region&module=RelatedCoverage&region=EndOfArticle&pgtype=article; Liz Robbins, *New York Can Destroy Documents, Judge Rules in Municipal ID Case*, N.Y. TIMES (Apr. 7, 2017), https://www.nytimes.com/2017/04/07/nyregion/new-york-can-destroy-documents-judge-rules-in-municipal-id-case.html; Ross Barkan, *What Happens to New York's Municipal ID Card Under the Trump*

Model Open Data Benefit Risk Analysis

In the open data context, considering only the risks of the dataset is merely one part of a balanced value equation; decision-makers must also take count of the project's benefits in order to make a final determination about whether to proceed with publishing the dataset openly.[22] For the purposes of this draft report, FPF developed a Model Open Data Benefit Risk Analysis based on risk assessment and de-identification frameworks developed by the National Institute of Standards and Technology and also builds on parallel efforts by researchers at the University of Washington, the Berkman Klein Center, and the City of San Francisco to develop robust risk-based frameworks for government data releases.[23] This template provides a structure for vetting potential open datasets in five steps:

Step 1: Evaluate the Information Contained in the Dataset. This step includes identifying whether there are direct or indirect identifiers, sensitive attributes, or information that is difficult to de-identify present in the dataset; assessing how linkable the information might be to other datasets; and considering the context in which the data was obtained.

Step 2: Evaluate the Benefits Associated with Releasing the Dataset. This step considers the potential benefits and users of the dataset, and assesses the magnitude of the potential benefits against the likelihood of their occurring.

Step 3: Evaluate the Risks Associated with Releasing the Dataset. This step considers the potential privacy risks and negative users of the dataset, and assesses the magnitude of the potential risks against the likelihood of their occurring.

Step 4: Weigh the Benefits against the Risks of Releasing the Dataset. This step combines the overall scores from steps 2 and 3 to determine an appropriate method for releasing (or not releasing) the dataset. Recommendations include releasing as open data, in a limited access environment, or not publishing at the current time. This section also overviews common methods for reducing re-identification risk in terms of their privacy-protective, utility, and operational impacts.

Step 5: Evaluate Countervailing Factors. This step provides a final opportunity to document any countervailing factors that might justify releasing a dataset openly regardless of its privacy risk, such as when there is a compelling public interest in the information.

[22] *See infra* Appendix B.

[23] *See* Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1968 (2015); Jan Whittington et al., *Push, Pull, and Spill: A Transdisciplinary Case Study in Municipal Open Government*, 30 BERKELEY TECH. L.J. 1899 (2015); Ben Green et al., *Open Data Privacy*, BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD (2017); DATASF, https://datasf.org/opendata/.

See Appendix B for the full template.

The City of Seattle as a Model Municipality

Given the risks described above, FPF conducted the following assessment to evaluate the City of Seattle as a model municipality based on its organizational structure and data handling practices related to open data. The assessment is grounded in public documentation and interviews with privacy, open data, and disclosure control experts and with Seattle IT and Open Data Leadership, departmental Open Data and Privacy Champions, and local community advisors.

Our scoring of the City of Seattle's practices in each of the following domains is based on the AICPA/CICA Privacy Maturity Model (PMM) levels, which reflect Generally Accepted Privacy Principles (GAPP):[24]

- Ad hoc – procedures or processes are generally informal, incomplete, and inconsistently applied.
- Repeatable – procedures or processes exist; however, they are not fully documented and do not cover all relevant aspects.
- Defined – procedures and processes are fully documented and implemented, and cover all relevant aspects.
- Managed – reviews are conducted to assess the effectiveness of the controls in place.
- Optimized – regular review and feedback are used to ensure continuous improvement towards optimization of the given process.

A key principle of the PMM approach is the recognition that "each organization's personal information privacy practices may be at various levels, whether due to legislative requirements, corporate policies or the status of the organization's privacy initiatives. It was also recognized that based on an organization's approach to risk, not all privacy initiatives would need to reach the highest level on the maturity model."[25]

Privacy leadership and program management

- Does the municipality employ a comprehensive, strategic, agency-wide privacy program regarding its open data initiatives?
- Has the municipality designated a privacy governance leader for Open Data?
- Is the Open Data program guided by core privacy principles and policies?
- Does the open data workforce receive effective privacy training and education?

[24] *See* AICPA/CICA PRIVACY TASK FORCE, AICPA/CICA PRIVACY MATURITY MODEL, (2011), https://www.kscpa.org/writable/files/AICPADocuments/10-229_aicpa_cica_privacy_maturity_model_finalebook.pdf

[25] *See id.*

- Are the municipality's open data privacy policies and procedures updated in light of ongoing monitoring and periodic assessments?

Benefit-risk assessment

- Does the Open Data program conduct a benefit-risk assessment to manage privacy risk in each dataset considered for publication?
- Are datasets assessed based on the identifiability, sensitivity, and utility of the data prior to release?
- Are inventories of published personally identifiable information maintained?
- Are benefit-risk assessments documented and regularly reviewed?
- Does the Open Data program have a mechanism in place to trigger re-assessment of a published dataset in light of new facts?

De-identification tools and strategies

- Does the Open Data program utilize technical, legal, and administrative safeguards to reduce re-identification risk?
- Does the Open Data program have access to disclosure control experts to evaluate re-identification risk?
- Does the Open Data program have access to appropriate tools to de-identify unstructured or dynamic data types? (e.g., geographic, video, audio, free text, real time sensor data).
- Does the Open Data program have policies and procedures for evaluating re-identification risk across databases? (e.g., risk created by intersection of multiple municipal databases, King County open data, Washington State open data, federal open data, commercial databases).
- Does the Open Data program evaluate privacy risk in light of relevant public records laws?

Data quality

- Does the municipality employ policies and procedures for the open data program to ensure that personally identifiable information is accurate, complete, and current?
- Does the Open Data program check for, and correct as appropriate, inaccurate, or outdated personally identifiable information?
- Are there procedures or mechanisms for individuals to submit correction requests for potentially incorrect data posted on the open data program?

Equity and fairness

- Were the conditions under which the data was collected fair? (e.g., were citizens aware that the data would be published on the open data portal? If data was acquired from a third party, were terms and conditions observed in the collection, use, maintenance, and sharing of the data?).

- Does the Open Data program assess the representativeness of the open data portal? (e.g. whether underserved or vulnerable populations are appropriately represented in the data, or whether underserved or vulnerable populations' interests are taken into account when determining what data to publish).
- Are any procedures and mechanisms in place for people to submit complaints about the use of data or about the open data process generally, as well as procedures for responding to those complaints?

Transparency and public engagement

- Does the Open Data program engage and educate the public about the benefits of open data?
- Does the Open Data program engage and educate the public about the privacy risks of open data?
- Does the Open Data program provide opportunities for public input and feedback about the program, the data available, and privacy, utility, or other concerns?
- Does the Open Data program engage with the public when developing of open data privacy protections?
- Does the Open Data program consider the public interest in determining what datasets to publish?
- Does the Open Data program communicate with the public about why some datasets may include personally identifiable information?

Model Open Data Risk Analysis applied to the current Seattle Open Data content

FPF will review a subset of content available on the open data program from high-risk agencies, as well as a random sample of additional agencies or datasets, and apply the final model template to evaluate their potential privacy risk relative to their potential benefits to the public. The datasets contemplated to be included in the Final Report are:

1. Real Time Fire 911 Calls
2. Building Permits (Current)
3. Sold Fleet Equipment
4. Seattle Communities Online Inventory
5. Road Weather Information Stations

Recommendations and Conclusion

As the City of Seattle Open Data program evolves and matures, it must continue developing the specialized resources and tools to address the privacy risks inherent in open data. The Seattle Open Data program will be building on a strong foundation, but there are always steps that can be taken to improve the depth and breadth of municipal privacy protections. The final report will detail concrete technical, operational, and organizational recommendations to elevate the Seattle Open Data program's approach to identifying and addressing privacy risks.

The City of Seattle is one of the most innovative cities in the country, with engaged and civic-minded citizenry, active city leadership, and technologically sophisticated business community. By continuing to complement its growing open data program with robust privacy protections and policies, it will be possible for the City of Seattle to live up to the promise of its Open Data Policy, supporting civic innovation while protecting individual privacy.

Appendix A: Additional Resources

AICPA/CICA PRIVACY TASK FORCE, AICPA/CICA PRIVACY MATURITY MODEL, (2011), https://www.kscpa.org/writable/files/AICPADocuments/10-229_aicpa_cica_privacy_maturity_model_finalebook.pdf.

Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1968 (2015), https://cyber.harvard.edu/publications/2016/Privacy_Aware_Government_Data_Releases.

SEAN BROOKS ET AL., AN INTRODUCTION TO PRIVACY ENGINEERING AND RISK MANAGEMENT IN FEDERAL SYSTEMS NISTIR 8062 (NIST Jan. 2017), http://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf.

JOSEPH A. CANNATACI, REPORT OF THE SPECIAL RAPPORTEUR ON THE RIGHT TO PRIVACY (Appendix on Privacy, Big Data, and Open Data) (Human Rights Council, Mar. 8, 2016), www.ohchr.org/Documents/Issues/Privacy/A-HRC-31-64.doc.

Lorrie Cranor, *Open Police Data Re-identification Risks*, TECH@FTC BLOG (April 27, 2016, 3:31 PM), https://www.ftc.gov/news-events/blogs/techftc/2016/04/ open-police-data-re-identification-risks

David Doyle, Open Government Data: an analysis of the potential impacts of an Open Data law for Washington State (2015) (unpublished M.P.P. thesis, University of Washington Bothell), https://digital.lib.washington.edu/researchworks/bitstream/handle/1773/34826/Doyle%20-%20Capstone.pdf?sequence=1.

Khaled El Emam, *A de-identification protocol for open data*, IAPP (May 16, 2016), https://iapp.org/news/a/a-de-identification-protocol-for-open-data/.

KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (CRC Press, 2013).

KHALED EL EMAM & WAËL HASSAN, A PRIVACY ANALYTICS WHITE PAPER: THE DE-IDENTIFICATION MATURITY MODEL (PrivacyAnalytics, 2013).

Federal Committee on Statistical Methodology, *Report on Statistical Disclosure Limitation Methodology* (Federal Committee on Statistical Methodology, Statistical Policy Working Paper No. 22, 2005), https://www.hhs.gov/sites/default/files/spwp22.pdf.

Kelsey Finch & Omer Tene, *Welcome to the Metropticon: Protecting Privacy in a Hyperconnected Town*, 41 FORDHAM URB. L.J. 1581 (2015).

Kelsey Finch & Omer Tene, *The City as a Platform: Enhancing Privacy and Transparency in Smart Communities*, CAMBRIDGE HANDBOOK OF CONSUMER PRIVACY (forthcoming).

ERICA FINKEL, DATASF: OPEN DATA RELEASE TOOLKIT (2016), https://drive.google.com/file/d/0B0jc1tmJAlTcR0RMV01PM2NyNDA/view.

SIMSON L. GARFINKEL, SP 800-188: DE-IDENTIFYING GOVERNMENT DATASETS (NIST draft. Aug. 2016), http://csrc.nist.gov/publications/drafts/800-188/sp800_188_draft2.pdf.

SIMSON L. GARFINKEL, NISTIR 8053: DE-IDENTIFYING PERSONAL INFORMATION (NIST Oct. 2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf.

Ben Green et al., *Open Data Privacy*, BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD (2017), https://dash.harvard.edu/bitstream/handle/1/30340010/OpenDataPrivacy.pdf.

Emily Hamilton, The Benefits and Risks of Policymakers' Use of Smart City Technology (Oct. 2016) (unpublished paper) (on file with the Mercatus Center at George Mason University).

INFORMATION COMMISSIONER'S OFFICE, ANONYMISATION: MANAGING DATA PROTECTION RISK (2012).

ISO/IEC CD 20889: Information technology – Security techniques – Privacy enhancing data de-identification techniques, https://www.iso.org/standard/69373.html?browse=tc.

ANNA JOHNSTON, DEMYSTIFYING DE-IDENTIFICATION: AN INTRODUCTORY GUIDE FOR PRIVACY OFFICERS, LAWYERS, RISK MANAGERS AND ANYONE ELSE WHO FEELS A BIT BEWILDERED, (Salinger Privacy, Feb. 2017).

JOINT TASK FORCE TRANSFORMATION INITIATIVE INTERAGENCY WORKING GROUP, GUIDE FOR CONDUCTING RISK ASSESSMENTS NIST 800-30 (NIST Sep. 2012), http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf .

Jeff Jonas & Jim Harper, *Open Government: The Privacy Imperative*, *in* OPEN GOVERNMENT: COLLABORATION, TRANSPARENCY, AND PARTICIPATION IN PRACTICE (O'Reilly Media, 2010).

ROB KITCHIN, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES (Sage, 1st ed. 2014).

YVES-ALEXANDRE DE MONTJOYE ET AL., UNIQUE IN THE CROWD: THE PRIVACY BOUNDS OF HUMAN MOBILITY (Scientific Reports 3, Mar. 25, 2013), https://www.nature.com/articles/srep01376.

SEAN A. MUNSON ET AL., ATTITUDES TOWARD ONLINE AVAILABILITY OF US PUBLIC RECORDS (2011).

Arvind Narayanan et al., *A Precautionary Approach to Big Data Privacy*, *in* 24 DATA PROTECTION ON THE MOVE: LAW, GOVERNANCE AND TECHNOLOGY SERIES (Serge Gutwirth, Ronald Leenes, Paul de Hert eds., 2016).

Opinion of the Article 29 Data Protection Working Party on Anonymisation Techniques, 2014.

Jules Polonetsky, Omer Tene & Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identificiton*, 56 SANTA CLARA L. REV. 594 (2016).

PRESIDENT'S COUNCIL OF ADVISORS ON SCIENCE AND TECHNOLOGY, EXEC. OFFICE OF THE PRESIDENT, Report to the President: Technology and the Future of Cities (Feb. 2016).

Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L REV. 703 (2016), http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1589/91WLR0703.pdf?sequence=1&isAllowed=y.

Sander v. State Bar of California, 58 Cal. 4th 300 (2013).

Jan Whittington et al., *Push, Pull, and Spill: A Transdisciplinary Case Study in Municipal Open Government*, 30 BERKELEY TECH. L.J. 1899 (2015), http://btlj.org/data/articles2015/vol30/30_3/1899-1966%20Whittington.pdf.

Alexandra Wood et al., *Privacy and Open Data Research Briefing*, BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD (2016), https://dash.harvard.edu/bitstream/handle/1/28552574/04OpenData.pdf?sequence=1.

Frederik Zuiderveen Borgesius et al., *Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework*, 30 BERKELEY TECH. L.J. 2075 (2015), http://btlj.org/data/articles2015/vol30/30_3/2073-2132%20Borgesius.pdf.

Seattle Resources

CITY OF SEATTLE, CITY OF SEATTLE 2017 OPEN DATA PLAN, http://www.seattle.gov/Documents/Departments/SeattleIT/City%20of%20Seattle%202017%20Open%20Data%20Plan.pdf.

CITY OF SEATTLE, OPEN DATA PLAYBOOK V. 1.0, http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPlaybook_Published_2016.08.pdf.

CITY OF SEATTLE, OPEN DATA POLICY, http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPolicyV1.pdf

CITY OF SEATTLE, OPEN DATA PROGRAM 2016 ANNUAL REPORT, https://www.seattle.gov/Documents/Departments/SeattleIT/Open%20Data%20Program%202016%20Annual%20Report.pdf.

CITY OF SEATTLE, PRIVACY PRINCIPLES, https://www.seattle.gov/Documents/Departments/InformationTechnology/City-of-Seattle-Privacy-Principles-FINAL.pdf.

Seattle Information Technology: Community Technology Advisory Board (CTAB), SEATTLE.GOV, https://www.seattle.gov/tech/opportunities/ctab.

Seattle Information Technology: Privacy, SEATTLE.GOV, http://www.seattle.gov/tech/initiatives/privacy.

Seattle Information Technology: Open Dataset Inventory – Privacy and PII, SEATTLE.GOV, https://view.officeapps.live.com/op/view.aspx?src=http://www.seattle.gov/Documents/Departments/SeattleIT/OpenDatasetInventory_Privacy_PII.docx.

Appendix B: Model Open Data Benefit Risk Analysis

Dataset: _____

Step 1: Evaluate the Information the Dataset Contains

Consider the following categories of information:

- *Direct Identifiers:* These are data points that identify a person without additional information or by linking to information in the public domain. "Personally Identifiable Information," or PII, often falls within this category. For example, they can be names, social security numbers, or an employee ID number. *See* PII/Privacy in the Open Dataset Inventory guidance. Publishing direct identifiers creates a *very high* risk to privacy because they directly identify an individual and can be used to link other information to that individual.

- *Indirect Identifiers:* These are data points that do not directly identify a person, but that in combination can single out an individual. This could include information such as birth dates, ZIP codes, gender, race, or ethnicity. In general, to preserve privacy, experts recommend including no more than 6-8 indirect identifiers in a single dataset.[26] If a dataset includes 9 or more indirect identifiers there is a *high* or *very high* risk to privacy because they can indirectly identify an individual.

- *Non-Identifiable Information:* This is information that cannot reasonably identify an individual, even in combination. For example, this might include city vehicle inventory, GIS data, or atmospheric readings. This data creates *very low* or *low* risk to privacy.

- *Sensitive Attributes:* These data points that may be sensitive in nature. Direct and indirect identifiers can be sensitive or not, depending on context. For example, this might include financial information, health conditions, or a criminal justice records. Sensitive attributes typically create *moderate*, *high*, or *very high* risk to privacy.

- *Spatial Data and Other Information that Is Difficult to De-identify:* Certain categories or data are particularly difficult to remove identifying or identifiable information from, including: geographic locations, unstructured text or free-form fields, biometric information, and photographs or videos.[27] If direct or indirect identifiers are in one of these data formats, they may create a *moderate*, *high*, or *very high* risk to privacy.

Consider how linkable the information in this dataset is to other datasets:

[26] *See* Khaled El Emam, *A De-Identification Protocol for Open Data*, IAPP (MAY 16, 2016), https://iapp.org/news/a/a-de-identification-protocol-for-open-data/.
[27] *See* GARFINKEL, *supra* note 8, at 32-33.

- Do any of the dataset's direct or indirect identifiers currently appear in other readily accessible open datasets, such as Data.Seattle.gov, Data.KingCounty.gov, or Data.WA.gov? If this information is present in multiple open datasets, it increases the chances of identifying an individual and increases the risk to privacy.

- How often is the dataset updated? In general, the more frequently a dataset is updated—every fifteen minutes versus every quarter, for example—the easier it is to re-identify an individual and the greater the risk to privacy.

- How often is the information in this dataset requested by public records.

Consider how the information in this dataset was obtained:

- In what context was this data collected? Is this data collected under a regulatory regime? Are there any conditions, such as a privacy policy or contractual term, attached to the data? If the personal information in this dataset collected directly from the individual or from a third party?

- Would there be a reasonable expectation of privacy in the context of the data collection? For example, if the public has no notice of the data collection or data are collected from private spaces, there may be an expectation of privacy.

- Was the collection of the information in this dataset controversial? Was any of the information in this dataset collected by surveillance technologies (e.g., bodyworn cameras, surveillance cameras, unmanned aerial vehicles, automatic license plate readers, etc.)?

- Has this dataset been checked for accuracy? Is there a mechanism for individuals to have information about themselves in this dataset corrected or deleted?

- Is there a concern that releasing this data may lead to public backlash or negative perceptions?

Step 2: Evaluate the Benefits Associated with Releasing the Dataset

List some of the foreseeable benefits of publishing the data fields included in this dataset. For example, measuring atmospheric data at particular locations over time may reveal useful weather patterns, and tracking building permit applications may reveal emerging demographic or commercial trends in particular neighborhoods.

Consider the likely users of this dataset. Who are the ideal users?

- ☐ Individuals
- ☐ Community Groups
- ☐ Journalists
- ☐ Researchers
- ☐ Companies or Private Entities
- ☐ Other Government Agencies or Groups
- ☐ Other: _____

Assess the scope of the foreseeable benefits of publishing the dataset on a scale of 1-10:

Qualitative Value	Quantitative Value	Description
Very High	10	The dataset will likely have *multiple compelling and important* utilities for individuals, the community, other organizations, or society.
High	8	The dataset will likely have a *compelling and important* utility for individuals, the community, other organizations, or society.
Moderate	5	The dataset will likely have a *clear* utility for individuals, the community, other organizations, or society. While the utility is clear, it is not as urgent as a "high" value.
Low	2	The dataset will likely have a *limited* utility for individuals, the community, other organizations, or society.
Very Low	0	The dataset will likely have *negligible* utility for organizations, the community, other organizations, or society.

Next, assess the likelihood that the desired benefits of releasing this dataset would occur:

Qualitative Value	Quantitative Value	Description
Very High	10	The benefit is *almost certain* to occur.
High	8	The benefit is *highly likely* to occur.
Moderate	5	The benefit is *somewhat likely* to occur.
Low	2	The benefit is *unlikely* to occur.
Very Low	0	The benefit is *highly unlikely* to occur.

Combining your rating of the foreseeable benefits of the dataset with the likelihood that these benefits will occur, assess the overall benefit of this dataset:

Likelihood of Occurrence	Impact of Foreseeable Benefits				
	Very Low Impact	Low Impact	Moderate Impact	High Impact	Very High Impact
Very High Likelihood	Low Benefit	Moderate Benefit	High Benefit	Very High Benefit	Very High Benefit
High Likelihood	Low Benefit	Moderate Benefit	Moderate Benefit	High Benefit	Very High Benefit
Moderate Likelihood	Low Benefit	Low Benefit	Moderate Benefit	Moderate Benefit	High Benefit
Low Likelihood	Very Low Benefit	Low Benefit	Low Benefit	Moderate Benefit	Moderate Benefit
Very Low Likelihood	Very Low Benefit	Very Low Benefit	Low Benefit	Low Benefit	Low Benefit

Step 3: Evaluate the Risks Associated with Releasing the Dataset

Consider the foreseeable privacy risks of this dataset[28]:

- *Re-identification (and false re-identification) impacts on individuals*
  - Would a re-identification attack on this dataset expose the person to identity theft, discrimination, or abuse?
  - Would a re-identification attack on this dataset reveal location information that could lend itself to burglary, property crime, or assault?
  - Would a re-identification attack on this dataset expose the person to financial harms or loss of economic opportunity?
  - Would a re-identification attack on this dataset reveal non-public information that could lead to embarrassment or psychological harm?

[28] Special thanks to Simson Garfinkel and Khaled El Emam whose works provide a foundation for articulating this analytic framework. *See* DE-IDENTIFICATION OF PERSONAL INFORMATION 32-33 (NIST 2015), DE-IDENTIFYING GOVERNMENT DATASETS SP 800-188; Khaled El Emam, *A De-Identification Protocol for Open Data*, IAPP (MAY 16, 2016), https://iapp.org/news/a/a-de-identification-protocol-for-open-data/; KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (2013).

- *Re-identification (and false re-identification) impacts on the organization*
  - Would a re-identification attack on this dataset lead to embarrassment or reputational damage to the City of Seattle?
  - Would a re-identification attack on this dataset harm city operations relying on maintaining data confidentiality?
  - Would a re-identification attack on this dataset expose the city to financial impact from lawsuits, or civil or criminal sanctions?
  - Would a re-identification attack on this dataset undermine public trust in the government, leading to individuals refusing to consent to data collection or providing false data in the future?

- *Data quality impacts*
  - Will inaccurate or incomplete information in this dataset create or reinforce biases towards or against particular groups?
  - Does this dataset contain any incomplete or inaccurate data that, if relied upon, would foreseeably result in adverse or discriminatory impacts on individuals?
  - Will any group or community's data be disproportionately included in or excluded from this dataset?
  - If this dataset is de-identified through statistical disclosure measures, did that process introduce significant inaccuracies or biases into the dataset?

- *Public impacts*
  - Does this dataset have information that would lead to public backlash if made public?
  - Will local individuals or communities be shocked or surprised by the information about themselves in this dataset?
  - Is it likely that the information in this dataset will lead to a chilling effect on individual, commercial, or community activities?
  - Is there any information contained within the dataset that would, if made public, reveal nonpublic information about an agency's operations?

Consider who could use this information improperly or in an unintended manner (including to re-identify individuals in the dataset):

- ☐ General public (individuals who might combine this data with other public information)
- ☐ Re-identification expert (a computer scientist skilled in de-identification)
- ☐ Insiders (a City employee or contractor with background information about the dataset)
- ☐ Information brokers (an organization that systematically collects and combines identified and de-identified

information, often for sale or reuse internally)

- ☐ "Nosy neighbors" (someone with personal knowledge of an individual in the dataset who can identify that individual based on the prior knowledge)
- ☐ Other: _____

Assess the scope of the foreseeable privacy risks of publishing the dataset on a scale of 1-10:

Qualitative Value	Quantitative Value	Description
Very High	10	The dataset will likely have *multiple severe or catastrophic* adverse effects on individuals, the community, other organizations, or society.
High	8	The dataset will likely have a *severe or catastrophic* adverse effect on individuals, the community, other organizations, or society.
Moderate	5	The dataset will likely have a *serious* adverse effect on individuals, the community, other organizations, or society.
Low	2	The dataset will likely have a *limited* adverse impact on individuals, the community, other organizations, or society,
Very Low	0	The dataset will likely have a *negligible* adverse impact on individuals, the community, other organizations, or society.

Next, assess the likelihood that the foreseeable privacy risks of releasing this dataset would occur:

Qualitative Value	Quantitative Value	Description
Very High	10	The risk is *almost certain* to occur.
High	8	The risk is *highly likely* to occur.
Moderate	5	The risk is *somewhat likely* to occur.
Low	2	The risk is *unlikely* to occur.
Very Low	0	The risk is *highly unlikely* to occur.

Combining your rating of the foreseeable risks of the dataset with the likelihood that these risks will occur, assess the overall risk of this dataset:

Likelihood of Occurrence	**Impact of Foreseeable Risks**				
	Very Low Impact	**Low Impact**	**Moderate Impact**	**High Impact**	**Very High Impact**
Very High Likelihood	Low Risk	Moderate Risk	High Risk	Very High Risk	Very High Risk

High Likelihood	Low Risk	Moderate Risk	Moderate Risk	High Risk	Very High Risk
Moderate Likelihood	Low Risk	Low Risk	Moderate Risk	Moderate Risk	High Risk
Low Likelihood	Very Low Risk	Low Risk	Low Risk	Moderate Risk	Moderate Risk
Very Low Likelihood	Very Low Risk	Very Low Risk	Low Risk	Low Risk	Low Risk

Step 4: Weigh the Benefits against the Risks of Releasing the Dataset

Step 4A: Combine the overall scores from the benefit and risk analyses to determine the appropriate solution for how to treat the dataset.

Benefit	**Risks**				
	Very Low Risk	**Low Risk**	**Moderate Risk**	**High Risk**	**Very High Risk**
Very High Benefit	Open	Open	Limit Access	Additional Screening	Additional Screening
High Benefit	Open	Limit Access	Limit Access	Additional Screening	Additional Screening
Moderate Benefit	Limit Access	Limit Access	Additional Screening	Additional Screening	Do Not Publish
Low Benefit	Limit Access	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish
Very Low Benefit	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish	Do Not Publish

- *Open*: Releasing this dataset to the public presents low or very low privacy risk to individuals, or the potential benefits of the dataset substantially outweigh the potential privacy risks. If the combination of risks and benefits resulted in an "Open" selection in the light green band, consider mitigating the data to further lower the risk.

- *Limit Access*: Releasing this data would create a moderate privacy risk, or the potential benefits of the dataset do not outweigh the potential privacy risks. In order to protect the privacy of individuals, limit access to the dataset such as by attaching contractual/Terms of Service terms to the data prohibiting re-identification attempts.
- *Additional Screening:* Releasing this dataset would create significant privacy risks and the potential benefits do not outweigh the potential privacy risks. In order to protect the privacy of individuals, formal application and oversight mechanisms should be considered (e.g., an institutional review board, data use agreements, or a secure data enclave).
- *Do Not Publish*: Releasing this dataset poses a high or very high risk to individual's privacy or the potential privacy risks of the dataset significantly outweigh the potential benefits. This dataset should remain closed to the public, unless there are countervailing public policy reasons for publishing it.

If the above table results in an "Open" categorization, then record the final benefit-risk score and prepare to publish the dataset openly. If the above table does *not* result in an "Open" categorization, then proceed to Step 4B by applying appropriate de-identification controls to mitigate the privacy risks for this dataset. The de-identification methods described below will be appropriate for some datasets, but not for others. Consider the level of privacy risks you are willing to accept, the overall benefit of the dataset, and the operational resources available to mitigate re-identification risk. Note that the more invasive the de-identification technique, the greater the loss of utility will be in the data, but also the greater the privacy protection will be.

Technical Controls[29]

Method	Description	Privacy Impact	Utility Impact	Operational Costs
Suppression	Removing a data field or an individual record to prevent the identification of individuals in small groups or those with unique characteristics.	Removing the field removes the risk created by those fields, and lowers the likelihood of linking one dataset to another based on that information.	This approach removes all utility added by the suppressed field or record, and could skew the results or give false impressions	This is a relatively low-cost method of de-identification. Removing entire fields of data can be both a quick and relatively low-tech process.

[29] Special thanks to the Berkman Klein Center for Internet & Society at Harvard University whose work provides a foundation for this analytic framework. BEN GREEN ET AL, OPEN DATA PRIVACY (2017), https://dash.harvard.edu/handle/1/30340010; Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1968 (2015), https://cyber.harvard.edu/publications/2016/Privacy_Aware_Government_Data_Releases.

		Removing individual records can also effectively protect the privacy of those individuals. Suppression cannot guarantee absolute privacy, because there is always a chance that the remaining data can be re-identified using an auxiliary dataset.	about the underlying data.	When removing records one-by-one, particularly large datasets, there is a risk that some records may be overlooked.[30]
Generalization/Blurring	Reducing the precision of disclosed data to minimize the certainty of individual identification, such as by replacing precise data values with ranges or sets.	The more specific a data value is, the easier it will generally be to single out an individual. However, even relatively broad categories cannot guarantee absolute privacy, because there is always a chance that the remaining data can be re-identified using an auxiliary dataset.	Generalizing data fields can render data useless for more granular analysis, and may skew results slightly or give false impressions about the underlying data.	Generalizing data fields can be a quick and straightforward process for reducing the identifiability of particular fields after the initial thresholds are set. In order to determine the appropriate level of generalization for particular data types, additional research or expert consultation may be required.
Pseudonymization	Replacing direct identifiers with a pseudonym (such as a randomly	Pseudonymization removes the association between an	Pseudonymization can allow for information about an individual to be	Pseudonymization can appear relatively straightforward

[30] *See* Fitzpatrick, *supra* note 9.

	generated value, an encrypted identifier, or a statistical linkage key).	individual and their data, and replaces it with a less easily identifiable key, lowering but not eliminating the risk of re-identification. Pseudonymization can be reversed in many circumstances, and are often considered personally identifiable information by privacy and data protection authorities.	linked across multiple records, increasing its utility for a wide variety of purposes.	and cost-effective, however creating *irreversible* pseudonyms suitable for open data release can require significant effort.[31] Most successful re-identification attacks on openly released data have come from data that was inadequately pseudonymized.[32]
Aggregation	Summarizing the data across the population and then releasing a report based on those statistics.	Aggregating data can be an effective method for protecting privacy as there is no raw data directly tied to an individual, however experts recommend	Aggregation is more useful for examining the performance of a group or cohort. Because the raw data is not presented, it cannot be relied on to generate additional insights.	This method of de-identification requires slightly more expertise than simply removing fields or records. After an initial learning curve, the method can be implemented

[31] *See* GARFINKEL, *supra* note 8, at 17.

[32] *See* Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk,* 91 WASH. L REV. 703 (2016), http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1589/91WLR0703.pdf?sequence=1&isAllowed=y; Jules Polonetsky, Omer Tene & Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identificiton,* 56 SANTA CLARA L. REV. 594 (2016).

		minimum cell sizes of 5-10 records.[33]		without significant costs. Expert consultants or guidance from federal statistical agencies may provide guidance in setting minimum cell sizes or addressing particular data types.[34]
Perturbation	An expert adds "noise" to the dataset (such as swapping values from one record to another, or replacing one value with an artificial value), making it difficult to distinguish between legitimate values and the "noise."	The false data in the field makes re-identification much less likely to occur. The noise makes it difficult to determine if re-identification is associated with a specific individual.	Utility decreases as the amount of noise in the data increases. The proportionate amount of legitimate data is reduced as false data is added.	This is costly in that it requires an expert. The type of noise, as well as the amount to be added will have a drastic difference, and to ensure a retention in utility, it must be completed by an expert. However, research shows that "even relatively small perturbations to the data may make re-identification difficult or impossible."[35]
k-anonymity	A technique to measure and limit	Privacy protection is greater as the	As with the above controls, the	This is a costly, complex, and

[33] *See* Khaled El Emam, Comment Letter on Proposed Rule to Protect the Privacy of Customers of Broadband and Other Telecommunications Services; Khaled El Emam, *Protecting Privacy Using k-Anonymity,* 15 J. AM. MED. INFORMATICS ASS'N (2008).

[34] *Id.*

[35] *See* GARFINKEL*, supra* note 8, at 29.

	how many individuals in a dataset have the same combination of identifiers. K-anonymity suppresses or generalizes identifiers and perturbs outputs until a particular k-value is reached.	value of "k" increases. Experts recommend that the k-value for open data sets should be at least k=11 (that is, for every combination of identifiers in a dataset, there should be at least 11 equivalent records).[36]	negative impact on utility increases as k-value increases. In order to achieve k=11, significant portions of some datasets may need to be suppressed or generalized.	time-consuming method. An expert in de-identification and k-anonymity is necessary to ensure that the k-value is correct and will provide the desired level of protection and utility. Subsequent research has led to additional requirements for the diversity of sensitive attribute within k-anonymous datasets (l-diversity) and statistical relationship to the original data (t-closeness).[37]
Differential Privacy	A set of techniques to mathematically determine if the result of an analysis of a dataset is the same before and after the removal of a single data	These techniques increase privacy for all individuals in a dataset and provide mathematical guarantees against re-identification for a certain period of time.	As differential privacy techniques rely on introducing noise, they decrease the accuracy of analysis performed on the dataset.	This operation requires an expert to calculate the leakage threshold, the amount of noise to add, and other statistical nuances. It also requires an online query system to

[36] El Emam, *supra* note 25.

[37] *See* GARFINKEL, *supra* note 8, at 12.

Commented [A4]: Consider clarifying that differential privacy is not the only control in this list that decreases the accuracy of an analysis on the released data. The data protection techniques identified in this report, such as suppression, generalization, and k-anonymity, often add a large amount of noise to the results of an analysis on data that have been transformed using these techniques. The literature has shown that such techniques can lead to results that differ substantially from the actual data and make the data unsuitable for certain analyses. Although differential privacy adds noise to the results of a statistical analysis, the amount of noise that is added can be negligible for large datasets. In fact, in some circumstances, the results of differentially private analyses are virtually indistinguishable from non-private analyses. In addition, the amount of noise that is added can be carefully tuned based on the balance of privacy-utility desired by the person applying the technique.

On another note related to utility, datasets or data records that cities classify as too privacy-sensitive to share in microdata (individual level) formats can potentially be safely shared using a differentially private tool, enabling analyses that are not possible using traditional data sharing models.

Commented [A2]: Systems that adhere to strong formal privacy models like differential privacy provide protection that is robust not only to certain types of re-identification attacks (such as record linkage attacks leveraging publicly available information) but also to a wide range of potential attacks, including attacks that unknown at the time of deployment, and do not require the person applying the technique to anticipate particular modes of attack. Further, these guarantees are not limited to a "certain period of time."

The protection provided by differential privacy differs from that of traditional de-identification techniques, which are often designed to address a narrow class of attacks. For example, many techniques explicitly or implicitly rely on a notion of privacy that is limited to record linkage attacks and require the person applying the technique to identify "direct or indirect identifiers" that appear in public databases and can be used to re-identify a person in a de-identified dataset.

	record. Differentially private datasets achieve this by adding small bits of random noise, and rely on online query systems to reduce "leakage" of data that might enable re-identification	Differential privacy techniques relies on limiting the number of queries completed to prevent maintain a proven minimum privacy threshold (often known as the "privacy budget"). The more queries performed on a function, the more the total "leakage" increases. The leakage can never decrease, and there is an acceptable level of leakage that can occur before a privacy risk becomes likely and the dataset must be abandoned.	The level of utility in a differentially private dataset is also dependent upon the number of queries to be made in the dataset. Once the leakage threshold is hit, the dataset becomes useless. However, if the desired task can be accomplished under the leakage threshold, the dataset retains great utility with little risk to privacy.	be established. Therefore, it carries a higher operational cost than other methods of de-identification. Differential privacy is an active research area, but to date it has only been applied to a few operational system.[38]
Synthetic Data	A process in which seed data from an original dataset is used to create artificial data that has some of the statistical characteristics as the seed data.[39]	Synthetic datasets can make it very difficult and costly to map artificial records to actual people, and supports mathematical privacy guarantees	Synthetic data "can be confusing to the lay public," as they may contain artificial individuals who "appear quite similar to actual individuals in the	Synthetic databases may be confusing to both researchers and lay people, requiring additional efforts to educate data users about the

Commented [A5]: Again, differential privacy is not limited to interactive, or query-based, mechanisms. Non-interactive mechanisms (which are not subject to a limited number of queries) can be implemented.

Commented [A1]: Differential privacy is a formal mathematical definition of privacy, which provides a provable guarantee of privacy against a wide range of potential attacks. It is not a single tool, but rather a standard, which many tools have been devised to satisfy. Some differentially private tools utilize an interactive query-based mechanism, and others are non-interactive, i.e., enabling data or data summaries to be released and used.

Commented [A6]: Other techniques such as suppression, generalization, and k-anonymity also require data privacy expertise in order to be applied effectively.

It could be noted that differentially private tools for use by non-experts in privacy, computer science, and statistics are currently in development. See, e.g., Marco Gaboardi et al., PSI (Ψ):a Private data Sharing Interface, Working Paper (2016), https://arxiv.org/abs/1609.04340.

This report could recommend that cities request that open data portal contractors such as Socrata and OpenGov develop and implement differentially private tools for use by non-experts into their platforms.

Commented [A3]: As mentioned above, differential privacy is not limited to interactive, or query-based, mechanisms. Various techniques, both interactive and non-interactive, can be rigorously shown to satisfy this definition. Government agencies such as the Census Bureau and corporations such as Google, Apple, and Uber use differential privacy to provide strong privacy protection when sharing statistics. In particular, the Census Bureau makes data available using a non-interactive differentially private mechanism. Additional tools
for differentially private analysis, including tools that are broadly-applicable and can be integrated with a wide range of existing software platforms, are under development at a number of research institutions.

[38] *See* GARFINKEL, *supra* note 8, at 7-9.
[39] GARFINKEL, *supra* note 11, at 48-49.

	Datasets may be partially synthetic (in which some of the data is inconsistent with the original dataset) or fully synthetic (in which there is no one-to-one mapping between any record in the original dataset and the synthetic dataset).[40]	with differential privacy that can remain in force "even if there are future data releases."[41]	population."[42] The utility of synthetic data also depends on the model used to create it. Synthetic databases do not need to be released via interactive query systems, as "the privacy loss budget can be spent in creating the synthetic dataset, rather than in responding to interactive queries."[43]	dataset's contents and limitations.

Commented [A7]: It could be clarified that this language is referring to differential privacy. Synthetic data is an example of a type of non-interactive data sharing model that can be designed to satisfy differential privacy. This example could also be mentioned in the discussion of differential privacy above.

Administrative and Legal Controls

Method	Description	Privacy Impact	Utility Impact	Operational Costs
Contractual provisions	Data is made available to qualified users under legally binding contractual terms, such as commitments not to attempt to re-identify individuals or link datasets, to	Contractual controls alone do not necessarily reduce the risk of re-identification, but when complementing the technical controls above can provide more flexible and	Contractual provisions do not impede utility for acceptable data uses, although the compliance costs may deter some potential data users.	Consistent contractual provisions must be developed and deployed, but this is a less extensive process than many of the technical measures above. Contractual provisions can also

[40] *Id. at* 49-54.
[41] *Id*. at 51.
[42] *Id*.
[43] *Id*. at 52.

	keep data private and secure, or to only use data for specified purposes.	contextual privacy protections. Contractual terms are more robust when backed up by audit requirements and penalties for noncompliance.		be tailored to the specific risk profiles of each dataset.
Data visualizations, contingency tables, summary statistics, etc.	Rather than providing users access to raw microdata, data may be presented in more privacy-protective formats, such as data visualizations (graphical depictions of a dataset's features or statistical properties), contingency tables (matrixes of the frequencies of certain variables), or summary statistics (particular aggregate measures of certain variables).	When data is released in non-tabular formats, individual data records are typically more obscure and harder to link to other auxiliary datasets, protecting individual privacy. On the other hand, some data display techniques, if not complemented by the technical controls above, may inadvertently draw attention to outliers (e.g., a data visualization may highlight unique values to a greater extent than a purely numerical publication).	Data released in these sorts of formats may still be highly useful for a range of purposes, although not all. These formats may also limit the ways in which datasets can be combined or built on to generate new insights. Visualizations and other alternative data formats may also be more engaging to the lay public than raw tabular data.	These are fairly low-cost approaches to limiting privacy risks, with numerous public resources readily available to open data program staff. Data that update frequently may be harder to maintain.
Access fees	Charging users for access to data	Because fees are likely to deter	The deterrent effect of access	Introducing access fees comes with

Commented [A8]: It's not clear why these controls are grouped together as a single item and why they are categorized as administrative and legal rather than technical.

Commented [A9]: It could also be noted that these data sharing models can be designed to satisfy formal privacy guarantees such as differential privacy.

Commented [A10]: It's not clear why this is identified as a privacy concern. Making it easier for members of the public to analyze the data provided through an open data portal would seem to be a feature, not a bug. If the discovery of outliers in the data would reveal privacy-sensitive information, this is an indication that more robust privacy protections are needed for the data release, as outliers could easily be found in a "purely numerical"/microdata release as well.

	increases accountability and may discourage improper use of data.	many casual browsers of a particular datasets, the likelihood of accidental re-identification of an individual by a curious friend, neighbor, or acquaintance generally decreases. Tiered fee structures (e.g., that charge more for commercial access or remote versus in-person data access) may also lower the risk of re-identification by other actors. Charging fees may also introduce registration and audit capabilities, allowing open data program staff to identify which data users accessed which datasets.	fees on the general public will impede the potential utility of the dataset and could limit access by some marginalized or vulnerable communities (e.g., those without credit cards, technological sophistication, or new market entrants).	initial and ongoing administrative overhead, and requires thoughtful determination of when particular datasets or classes of users warrant the use of fees.
Data enclaves	Physical or virtual environments are created that enable "authorized users to access	Risks of re-identification are almost entirely removed by restricting external access to even de-	Data utility can be maximized for qualified researchers, as privacy protections are no	There are significant operational costs to maintaining a secure data enclave, including

	confidential data and analyze the data using provided statistical software."[44]	identified data and introducing accountability and oversight measures. Technical controls may not need to be as strict, when complemented by administrative and legal safeguards (such as requiring researchers to apply for access, describe the proposed research, agree to confidentiality laws and penalties, audit logs, and authentication measures).	longer purely technical. Researchers may be limited in what research questions can be asked and in the format of their results. But data utility is completely removed for any individual or organization that is not approved to access the dataset.	establishing policies and procedures for granting qualified researcher queries, for processing queries on de-identified data, for establishing the enclave, and for monitoring the program over time.
Ethical oversight/advisory review board	Particularly risk or ambiguous policy decisions about a dataset are escalated to an external advisory group with broad expertise and community engagement for further review.[45]	External review boards with diverse backgrounds and subject matter expertise can more robustly debate the benefits and risks of releasing a dataset and can address any	An external review board may determine that a dataset's utility ultimately outweighs its impact on individual privacy; it may also determine that the benefits do *not* outweigh the risks.	Establishing and maintaining a body of experts can be a challenging operational endeavor, although guidance and models from academic data

[44] *See* Micah Altman et al., *supra* note 22, at 40; GARFINKEL, *supra note* 11 at ix.

[45] *See generally* CONFERENCE PROCEEDINGS: BEYOND IRBS: ETHICAL GUIDELINES FOR BIG DATA RESEARCH, FUTURE OF PRIVACY FORUM (Dec. 10, 2015), https://fpf.org/wp-content/uploads/2017/01/Beyond-IRBs-Conference-Proceedings_12-20-16.pdf.

		additional dimensions not captured by the privacy risk assessment (e.g., ethical, scientific, or community factors).		research are available.[46]

Step 4B: After determining and applying appropriate privacy controls and mitigations for the dataset, re-assess the overall risks and benefits of the dataset (Steps 1-3). Note any mitigation steps taken, and record the final benefit-risk score:

Benefit	**Risks**				
	Very Low Risk	**Low Risk**	**Moderate Risk**	**High Risk**	**Very High Risk**
Very High Benefit	Open	Open	Limit Access	Additional Screening	Additional Screening
High Benefit	Open	Limit Access	Limit Access	Additional Screening	Additional Screening
Moderate Benefit	Limit Access	Limit Access	Additional Screening	Additional Screening	Do Not Publish
Low Benefit	Limit Access	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish
Very Low Benefit	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish	Do Not Publish

If the score is still not "Open," consider using another re-identification method. If this is not possible, then determine whether to publish the dataset. If the dataset is categorized as "Additional Screening" or "Do Not Publish" but there may be countervailing public policy factors that should be considered, move on to Step 5.

- *Open*: Releasing this dataset to the public presents low or very low privacy risk to individuals, or the potential benefits of the dataset substantially outweigh the potential privacy risks. If the

[46] *See* 45 C.F.R. 46.102; OMER TENE & JULES POLONETSKY, BEYOND IRBS: ETHICAL GUIDELINES FOR BIG DATA RESEARCH 1 (Dec. 2015), https://bigdata.fpf.org/wp-content/uploads/2015/12/Tene-Polonetsky-Beyond-IRBs-Ethical-Guidelines-for-Data-Research1.pdf.

combination of risks and benefits resulted in an "Open" selection in the light green band, consider mitigating the data to further lower the risk.

- *Limit Access*: Releasing this data would create a moderate privacy risk, or the potential benefits of the dataset do not outweigh the potential privacy risks. In order to protect the privacy of individuals, limit access to the dataset such as by attaching contractual/Terms of Service terms to the data prohibiting re-identification attempts.
- *Additional Screening:* Releasing this dataset would create significant privacy risks and the potential benefits do not outweigh the potential privacy risks. In order to protect the privacy of individuals, formal application and oversight mechanisms should be considered (e.g., an institutional review board, data use agreements, or a secure data enclave).
- *Do Not Publish*: Releasing this dataset poses a high or very high risk to individual's privacy or the potential privacy risks of the dataset significantly outweigh the potential benefits. This dataset should remain closed to the public, unless there are countervailing public policy reasons for publishing it.

Step 5: Evaluate Countervailing Factors

Sometimes, a dataset with a very high privacy risk is still worth releasing into the open data program in light of public policy considerations. For example, a dataset containing the names and salaries of elected officials would likely be considered high-risk due to the inclusion of a direct identifier. However, there is a compelling public interest in making this information available to citizens that outweighs the risk to individual privacy.

Additionally, there are always risks associated with maintaining and releasing any kind of data relating to individuals. Two key considerations when deciding whether to release the data irrespective of a potentially high or very high risk to individual privacy are:

- 1) If you are on the edge between two categories, analyze the dataset holistically but err on the side of caution. A dataset that is not released immediately can still be released at another date, as additional risk mitigation techniques become available. A dataset that has been released publicly, however, cannot ever be fully pulled back, even if it is later discovered to pose a greater risk to individual privacy. Be particularly cautious about moving data from an original recommendation of *Do Not Publish* to *Open*, and ensure that the potential benefits of releasing the data are truly so likely and compelling that they outweigh the existing privacy risks.
- 2) Any time you deviate from the original analysis, document your reasoning for doing so. This will not only help you decide whether the deviation is, in fact, the correct decision, but also provides accountability. Should the need arise, you will have a record of your reasoning, including analysis of the expected benefits and the recognized risks at the time. Where personally identifiable information is published notwithstanding the privacy risk, accountability mechanisms help maintain trust in the Open Data program that may otherwise be lost.

1400 EYE STREET NW | SUITE 450 | WASHINGTON, DC 20005 · **FPF.ORG**