



# DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

## Open Data Privacy

The Harvard community has made this article openly available.  
[Please share](#) how this access benefits you. Your story matters.

Citation	Green, Ben, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer, and Susan Crawford. 2017. Open Data Privacy (2017). Berkman Klein Center for Internet & Society Research Publication.
Published Version	<a href="https://cyber.harvard.edu/publications/2017/02/opendataprivacyplaybook">https://cyber.harvard.edu/publications/2017/02/opendataprivacyplaybook</a>
Accessed	October 25, 2017 6:13:38 PM EDT
Citable Link	<a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010">http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010</a>
Terms of Use	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a>

*(Article begins on next page)*



# OPEN DATA PRIVACY

A risk-benefit, process-oriented approach to sharing and protecting municipal data

Ben Green | Gabe Cunningham | Ariel Ekblaw | Paul Kominers | Andrew Linzer | Susan Crawford



**BERKMAN  
KLEIN CENTER**  
FOR INTERNET & SOCIETY  
AT HARVARD UNIVERSITY

**RESPONSIVE  
COMMUNITIES**  
digital justice < > data stewardship



## Acknowledgments

This report would not have been possible without valuable contributions from the following individuals:

Micah Altman, Brenna Berman, Beth Blauer, Joy Bonaguro, David Cruz, Matt Daley, David Eaves, David Edinger, Erica Finkle, Jascha Franklin-Hodge, Garlin Gilchrist, Mark Headd, Chad Kenney, Patrick Lamphere, Howard Lim, Amen Ra Mashariki, Michael Mattmiller, Yves-Alexandre de Montjoye, Andrew Nicklin, Vitaly Shmatikov, Nick Sinai, Maria Smith, Latanya Sweeney, Waide Warner, Mitch Weiss, Oliver Wise, Josh Wolff, and Alexandra Wood.



## Suggested Citation

Green, Ben, Gabe Cunningham, Ariel Ekblaw, Paul Kominers, Andrew Linzer, and Susan Crawford. Open Data Privacy (2017). Berkman Klein Center Research Publication. Available at DASH: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010>.

## Contact

Ben Green  
bgreen@g.harvard.edu  
Berkman Klein Center for Internet and Society  
23 Everett St, Cambridge MA 02138  
cyber.harvard.edu



This work is licensed under a Creative Commons Attribution 4.0 International License.

# EXECUTIVE SUMMARY

Cities today collect and store a wide range of data that may contain sensitive information about residents. As cities embrace open data initiatives, more of this information is released to the public. While opening data has many important benefits, sharing data comes with inherent risks to individual privacy: released data can reveal information about individuals that would otherwise not be public knowledge.

At the heart of this dilemma lie two traits of granular (i.e., multi-dimensional, raw, and record-level) open data:

- **Benefit** (utility): Because it enables varied and detailed analyses, granular data is the most interesting and useful for businesses, policymakers, researchers, and the public.
- **Risk** (privacy): Because it contains the most detailed information, granular data often includes personally sensitive information.

These two attributes are often in conflict because less granular data protects privacy but is less valuable as an asset to promote transparency, enable innovation, and aid research. Just as open data is not valuable unless it is detailed, opening data will not be effective if it necessarily involves risks to individual privacy. It is therefore critical to develop effective approaches to balance these benefits and risks, enabling cities to release open data without unduly compromising sensitive information.

Traditional privacy and anonymization frameworks focus on identifying and removing personally identifiable information (PII).<sup>1</sup> Recent research, however, has revealed that this framework is unsustainable and ineffective. Because so much data is now available from a wide variety of sources, and because databases can be manipulated and combined in complex and unpredictable ways, information that might not be deemed PII can lead to the identification of a specific individual and enable inferences to be made about that individual.

In 2014, the President's Council of Advisors on Science and Technology (PCAST) wrote, "By data mining and other kinds of analytics, non-obvious and sometimes private information can be derived from data that, at the time of their collection, seemed to raise no, or only manageable, privacy issues" and that "one can never know what information may later be extracted from any particular collection of big data."<sup>2</sup> A 2015 study of anonymity in metadata concludes, "Our results render the concept of PII, on which the applicability of U.S. and European Union (EU) privacy laws depend, inadequate."<sup>3</sup>

---

<sup>1</sup>Paul M Schwartz and Daniel J Solove, "Reconciling Personal Information in the United States and European Union," *California Law Review* 102, no. 4 (2014).

<sup>2</sup>President's Council of Advisors on Science and Technology. "Big Data and Privacy: A Technological Perspective." (2014) [https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_big\\_data\\_and\\_privacy\\_-\\_may\\_2014.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf).

<sup>3</sup>Yves-Alexandre de Montjoye et al., "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science* 347, no. 6221 (2015).

Unfortunately, there are few clearly defined sets of data attributes that do or do not reveal private information. Computer scientists Arvind Narayanan and Vitaly Shmatikov write,

“The versatility and power of re-identification algorithms imply that terms such as ‘personally identifiable’ and ‘quasi-identifier’ simply have no technical meaning. While some attributes may be uniquely identifying on their own, any attribute can be identifying in combination with others.”<sup>4</sup>

Neither regulations nor ordinances provide sufficient clarity, as data publishers and consumers are moving faster than lawmakers. As PCAST writes, existing laws based on outdated PII concepts may give a false sense of security: “[A]nonymization is already rooted in the law, sometimes giving a false expectation of privacy where data lacking certain identifiers are deemed not to be personally identifiable information and therefore not covered by such laws as the Family Educational Rights and Privacy Act (FERPA).”<sup>5</sup> Thus, while ensuring legal compliance is a natural starting point for crafting data release policies, cities must look beyond legal compliance when crafting data release procedures and strategies.

This leaves open data officials in the position of often serving as de facto privacy arbiters. Because there is no clear consensus regarding how to add privacy protections prior to releasing datasets, municipal officials use varying processes to manage how data is collected, maintained, and released. While cities are eager to release data, they also want to avoid privacy mishaps that could emerge immediately or in the future and undermine an otherwise beneficial program.

Without the ability to manage and mitigate privacy risks in an effective and auditable manner, open data programs will be unable to fully realize the benefits stemming from collecting, using, and sharing data. Effective privacy management is essential to maximizing the impact of open data. As Marc Groman, Senior Adviser for Privacy at the Office of Management and Budget describes, “a well-resourced, well-functioning privacy program [...] will promote innovation [...] and enable more information sharing.”<sup>6</sup>

The goal of this document is to take a first step toward codifying responsible privacy-protective approaches and processes that could be adopted by cities and other groups that are publicly releasing data. Our report is organized around four recommendations.

---

<sup>4</sup>Arvind Narayanan and Vitaly Shmatikov, “Myths and fallacies of ‘Personally identifiable information,’” *Communications of the ACM* 53, no. 6 (2010).

<sup>5</sup>President’s Council of Advisors on Science and Technology, “Big Data and Privacy: A Technological Perspective.”

<sup>6</sup>Jill R. Aitoro, “Defining privacy protection by acknowledging what it’s not,” *Federal Times*, March 8, 2016 <http://www.federaltimes.com/story/government/interview/one-one/2016/03/08/defining-privacy-protection-acknowledging-what-s-not/81464556/>.

# RECOMMENDATIONS

## 1. Conduct risk-benefit analyses to inform the design and implementation of open data programs.

The complexity of balancing utility and privacy in open data means that there is no “correct” decision for any dataset: releasing data carries benefits for the public as well as potential risks to individual privacy. Cities have both legal and ethical obligations to protect the individuals whose sensitive information they possess. That risks in this area are inevitable, however, does not mean that cities should stop releasing data; cities will need to become comfortable with a certain level of risk. As cities move to release open data, they must become informed about the risks involved as well as the privacy and security controls available to mitigate these risks. Cities should then conduct risk-benefit analyses to evaluate whether the value that open datasets could yield outweighs the potential privacy risks of releasing that data.

## 2. Consider privacy at each stage of the data lifecycle.

Cities have traditionally focused on privacy only when releasing open data, but effective privacy management requires privacy to be taken into account at all stages of a dataset’s lifecycle. Privacy risks can emerge and be realized throughout the open data lifecycle of collection, maintenance, release, and deletion. Certain risks are best addressed at each stage. Cities should make privacy-aware decisions before they collect data, before they store or process data, and before and after they release data. Implementing appropriate safeguards at all stages is particularly important for municipal governments because public records requests can prompt the release of data at any time. Cities should therefore carefully consider what data they should collect and store, and not just what data they will release.

## 3. Develop operational structures and processes that codify privacy management widely throughout the City.

Because there is no one-size-fits-all solution to data privacy, cities should develop clear and consistent data management processes to continually evaluate the risks and benefits of releasing data. To this end, open data management should shift from output assessment (“have we released enough data?”) to process-oriented standards (“have we evaluated and acted upon the risks and benefits related to collecting, managing, and sharing this dataset?”). Privacy efforts should draw on the field of information security, which focuses on risk-mitigation processes. These policies should be specifically designed to reflect the priorities that the city intends to support, and to meet the city’s needs to comply with open records and privacy laws. Cities should be able to document the steps they have followed. Critical to these aims is institutionalizing privacy awareness through programs such as employee trainings that ensure privacy policies and priorities are understood widely. Similarly, a toolkit of data-sharing strategies that go beyond the binary open/closed distinction will help cities maximize the value their data provides. In order to ensure ongoing compliance within the rapidly-evolving data privacy ecosystem, cities should periodically review their practices and risk-benefit assessments.

#### 4. Emphasize public engagement and public priorities as essential aspects of data management programs.

A primary motivation for launching open data initiatives is that open data holds tremendous promise to improve government transparency and accountability. Yet open data is merely a means toward transparency and accountability, not an end in itself. Cities can further the goal of accountability by being transparent about their open data decisions: rather than focusing on publishing the most data, open data leaders should also evaluate their efforts based on the extent to which data are released in a transparent and accountable manner. When publishing new open data, for example, cities should share their rationale for making the data available, expected benefits of releasing that data weighed against the privacy risks, and measures that have been implemented to protect privacy. Ultimately, a successful open data program relies on public trust that the government is a responsible steward of individual data. Decisions regarding how to release data should therefore be made with meaningful consideration for the public's priorities regarding what information is released. Cities should also carry out proactive and ongoing engagement to incorporate public input into their open data decisions and to keep the public informed about new developments.



# OUTLINE

Each chapter of this report is dedicated to one of these four recommendations, and provides fundamental context along with specific suggestions to carry out these high-level recommendations.

[Chapter 1](#) introduces the concepts and practices behind risk-benefit analyses, along with background on the privacy risks of open data and the limits of existing de-identification approaches.

[Chapter 2](#) outlines a lifecycle approach to managing privacy in open data, describing a variety of steps that cities can take to better protect individual privacy.

[Chapter 3](#) emphasizes the importance of internal practices, focusing on the need to institutionalize effective privacy management at every stage of the data lifecycle and at all levels of the organization.

[Chapter 4](#) describes the role of public engagement, emphasizing the need for a nuanced understanding of public concerns and proactive engagement regarding open data decisions.

Each chapter includes the following components:

**Summary:** an overview of the key background and motivation for the recommendation.

**Background:** the context necessary to understand the motivation for the recommendation and its implementation.

**Take action:** a set of practices that can be implemented in accordance with the recommendation.

**In practice:** a case study highlighting why the recommendation is necessary or how an organization is implementing best practices.

Finally, the [Appendix](#) synthesizes key elements of the report into an Open Data Privacy Toolkit that cities can use to manage privacy when releasing data.



# TABLE OF CONTENTS

<b>1. Conduct risk-benefit analyses to inform the design and implementation of open data programs . . . . .</b>	<b>9</b>
1.1 Determine the desired benefits of releasing each element of open data . . . . .	13
1.2 Recognize the limits of de-identification techniques and evaluate the privacy risks of releasing data . . . . .	17
1.3 Consider a diversity of potential mitigations and choose the one best calibrated to the specific risks and benefits of the data . . . . .	26
<b>2. Consider privacy at each stage of the data lifecycle . . . . .</b>	<b>32</b>
2.1 Collect: Be mindful of privacy before collecting any data . . . . .	33
2.2 Maintain: Keep track of privacy risks in all data stored and maintained. . . . .	36
2.3 Release: Evaluate datasets for privacy risks and mitigate those risks before releasing data . . . . .	40
2.4 Delete: Where appropriate, retire data stored internally, turn off automatic collection, and remove data shared online to mitigate privacy risks that result from the accumulation of data . . . . .	44
<b>3. Develop operational structures and processes that codify privacy management throughout the open data program . . . . .</b>	<b>49</b>
3.1 Increase internal awareness of and attention to privacy risks . . . . .	51
3.2 Periodically audit data and processes to ensure privacy standards continue to be upheld . . . . .	55
3.3 Account for the unique risks and opportunities presented by public records laws . . . . .	58
3.4 Develop a portfolio of approaches for releasing and sharing data . . . . .	63
<b>4. Emphasize public engagement and public priorities as essential aspects of open data programs . . . . .</b>	<b>67</b>
4.1 Garner support for open data by sharing the benefits and successful uses of open data . . . . .	68
4.2 Develop constituency trust by considering public expectations and acting as responsible data stewards . . . . .	73
4.3 Bake public input into all aspects of the open data program . . . . .	80
4.4 Be transparent and accountable regarding all practices related to open data . . . . .	83
4.5 Build support for new initiatives before rolling them out . . . . .	87
<b>Conclusion . . . . .</b>	<b>90</b>
<b>References . . . . .</b>	<b>91</b>
<b>Open Data Privacy Toolkit . . . . .</b>	<b>97</b>

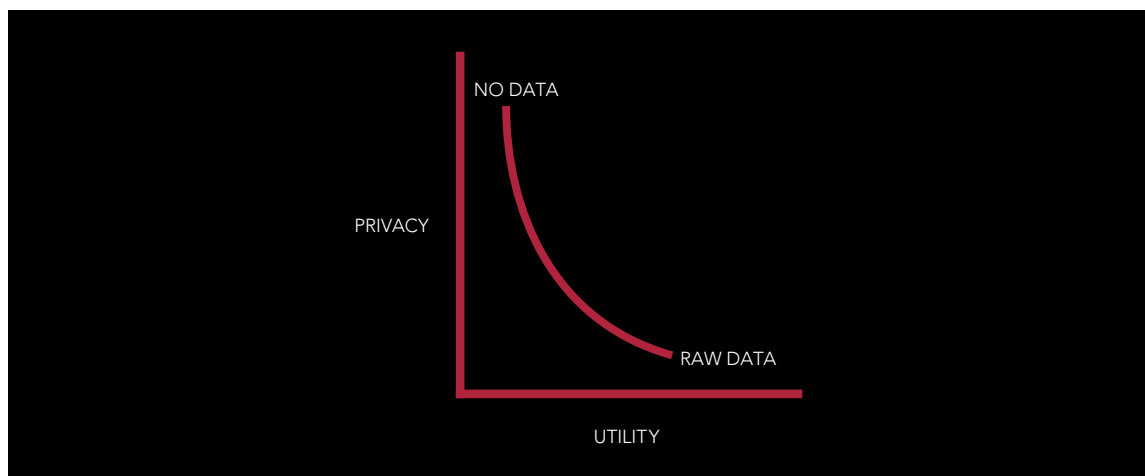
# 1. CONDUCT RISK-BENEFIT ANALYSES TO INFORM THE DESIGN OF OPEN DATA PROGRAMS.

Sharing data in any form involves a tradeoff between providing value and exposing private information (or information that could later be triangulated to reveal personal information). Risk-benefit analyses help cities find data management solutions that maximize benefits and minimize risks.

Opening access to municipal data involves a series of tradeoffs. Releasing data can increase transparency in government, allow citizens to engage with their cities, and empower entrepreneurs to build innovative tools and applications. As cities around the United States embrace open data practices, more of these benefits become possible every day. But because much of this data contains information related to individual citizens, releasing it also presents the potential for private information to be revealed. Limiting privacy risks when releasing data remains an unsolved challenge for open data initiatives.

It is difficult to balance the utility gained from allowing public consumption of data with the risks to individual privacy that occur when releasing that data. The broad tradeoff is clear: raw, granular data about people is the most useful but also the most sensitive; data with fewer fields or less-informative features better protects privacy but limits utility (*Figure 1*). And broad sections of the public may be benefited by releases of data that might be felt by a few people to invade their personal privacy. This puts open data programs in a difficult situation. Unfortunately, there are no clear boundaries to define when data is “useful” or “sensitive” — these two traits coexist and are often in conflict. That risks in this area are inevitable, however, does not mean that cities should stop releasing data; cities will need to become comfortable with a certain level of risk.

*Figure 1. The tradeoff between privacy and utility*



One example of information at the nexus of this multi-dimensional dilemma is crime data about sexual assault and domestic violence. Crime data is simultaneously one of the most useful and desired municipal datasets and, especially in the case of sexual assault data, one of the most sensitive. While open data about sexual assault and domestic violence can be a powerful tool for research and advocacy, the potential re-identification of victims can have significant consequences. A survey by the National Domestic Violence Hotline found that 60% of women who had experienced partner abuse and not contacted the police attributed their reticence to “not wanting police involvement due to a desire for privacy.”<sup>7</sup> Moreover, reports estimate that “an estimated 211,200 rapes and sexual assaults went unreported to police each year between 2006 and 2010.”<sup>8</sup> This represents 65% of such incidents. For a crime that already exists in the shadows, careless disclosure of victim identities could have chilling effects to an already under-reported crime. Given that re-identification of victims of domestic violence and sexual assault “could put their safety and security at risk”, FTC Chief Technologist Lorrie Cranor writes, “it is critical that we think through data re-identification issues before releasing data to the public.”<sup>9</sup>

To navigate this complexity, cities can conduct risk-benefit analyses, a flexible framework for weighing the advantages and disadvantages of any practice or process. For open data, such analyses can illuminate the specific features of datasets that contribute to their risks (release of sensitive information) and benefits (utility from releasing data).

On the federal level, numerous policies emphasize the need for privacy impact assessments of data before it is released. For example, the Open Data Policy of 2013 mandates agencies to take a risk-based approach to data privacy, writing, “The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified.”<sup>10</sup>

The National Institute of Standards and Technology (NIST) provides a useful framework for conducting risk assessments.<sup>11</sup> NIST’s process measures risk by identifying five components: vulnerabilities, threat events, threat sources, impact, and likelihood. These elements together define the level of risk; *Figure 2* on the following page displays how these components interact with one another to produce negative outcomes.

---

<sup>7</sup>The National Domestic Violence Hotline. “Who Will Help Me? Domestic Violence Survivors Speak Out About Law Enforcement Responses.” (2015) <http://www.thehotline.org/resources/law-enforcement-responses/>.

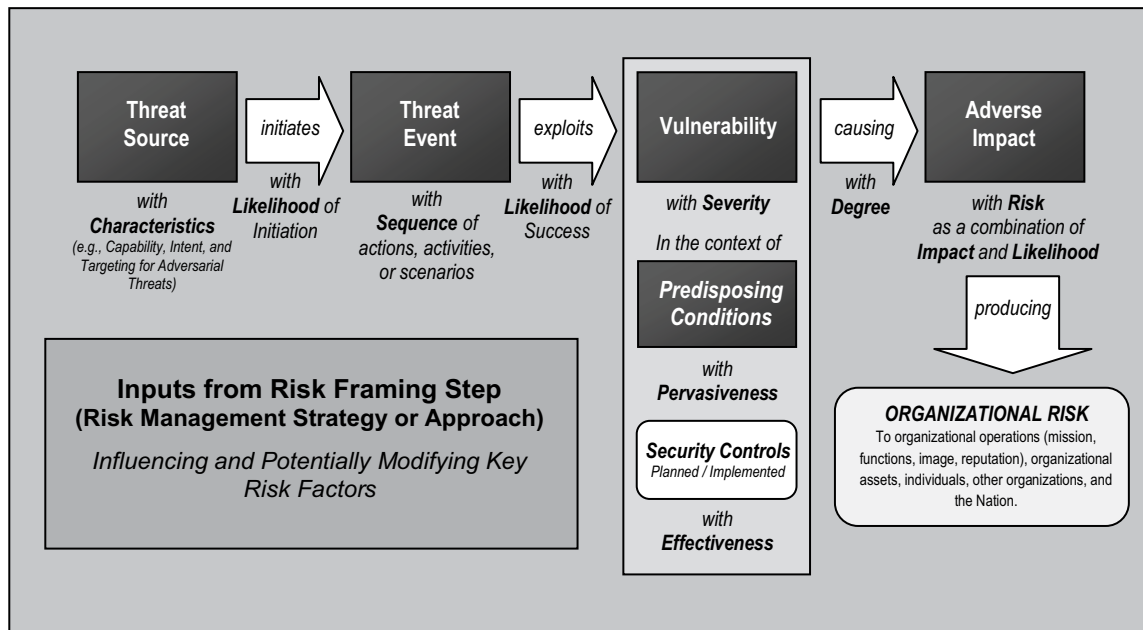
<sup>8</sup>Lynn Langton et al. “Victimizations Not Reported To The Police, 2006-2010.” Bureau of Justice Statistics (2012) <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=4962>.

<sup>9</sup>Lorrie Cranor, “Open Police Data Re-identification Risks,” <https://www.ftc.gov/news-events/blogs/techftc/2016/04/open-police-data-re-identification-risks>.

<sup>10</sup>Sylvia M Burwell et al., “Open Data Policy—Managing Information as an Asset,” Executive Office of the President, Office of Management and Budget 2013. <https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>.

<sup>11</sup>National Institute of Standards and Technology. “Guide for Conducting Risk Assessments.” (2012) <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf>.

Figure 2. Risk assessment overview



From National Institute of Standards and Technology. "Guide for Conducting Risk Assessments." (2012) <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf>.

In order to incorporate benefit into our open data risk-benefit framework, we have adapted these components for risk into an equivalent process to assess benefits. The table below provides definitions for the terms comprising risk and benefit.

	RISK	BENEFIT
<b>Data Attribute</b> (vulnerability, asset)	<i>Vulnerabilities</i> are attributes that increase an organization's susceptibility to negative outcomes (threat events).	<i>Assets</i> are attributes that increase an organization's ability to capture positive outcomes (advantage events).
<b>Event</b> (threat event, advantage event)	<i>Threat events</i> are negative outcomes that arise out of vulnerabilities.	<i>Advantage events</i> are positive outcomes that arise out of assets.
<b>Source</b> (threat source, advantage source)	<i>Threats sources</i> are people or organizations who seek to initiate threat events.	<i>Advantage sources</i> are people or organizations who seek to initiate advantage events.
<b>Likelihood</b>	<i>Likelihood</i> is the chance that a threat event occurs through a threat source successfully exploiting a vulnerability.	<i>Likelihood</i> is the chance that an advantage event occurs through an advantage source successfully exploiting an asset.



	RISK	BENEFIT
Impact	<i>Impact</i> is the negative effect caused by a threat event. This is a factor of both scale (the number of people affected) and severity (the damage caused to each person).	<i>Impact</i> is the positive effect caused by an advantage event. This is a factor of both scale (the number of people affected) and severity (the utility provided to each person).
Outcome (risk, benefit)	<i>Risk</i> is a synthesis of the likelihood and impact that describes the overall danger for an organization.	<i>Benefit</i> is a synthesis of the likelihood and impact that describes the overall opportunity for an organization.

Cities can combine these components to evaluate the benefits and risks of open data through an approach focused on data attributes. In a data-oriented approach, a city would assess risks by first identifying vulnerabilities in its data. For each of these vulnerabilities, the city would then list the threat events that may exploit these vulnerabilities and the threat sources that may initiate these threat events. Then a city would assign each threat event a likelihood of occurring and an impact if it were to occur. Finally, based on its determination of the likelihood and impact, the city can synthesize these components to determine overall risk. The same process, using the equivalent terms for benefits, would apply to evaluate the benefits of data.

The results of the risk and benefit assessments comprise a risk-benefit ratio, which compares the overall risks and benefits present in a dataset. A high risk-benefit ratio means that the risks are high relative to the benefits, while a low risk-benefit ratio means that the benefits outweigh the risks. The risk-benefit ratio can guide decisions regarding whether and how to release or withhold data.

If the risk-benefit ratio implies that there are more risks than benefits to releasing data, cities should attempt to improve this ratio through mitigations (interventions that decrease risk). A mitigation should explicitly target the vulnerabilities that lead to risks while attempting to maintain the assets that lead to benefits. A successful mitigation should lead to a more palatable risk-benefit ratio (typically by decreasing risk more than it decreases benefit).

To illuminate how the risk-benefit framework enables responsible data management, this chapter explains how to conduct a risk-benefit analysis for municipal open data.

# 1.1 DETERMINE THE DESIRED BENEFITS OF RELEASING EACH ELEMENT OF OPEN DATA.

Assessing the best way to release a dataset requires clear objectives for that data once released. Identifying useful data, potential users, and desired outcomes is essential for evaluating the potential positive impacts of open data.

Open data programs outline high-level objectives such as increasing transparency, improving internal efficiency, stimulating economic growth, and improving quality of life for residents. Nevertheless, these same initiatives often treat all datasets equally and evaluate progress based on the number of datasets released. In reality, each dataset contributes a different amount to these goals.

Making an educated decision about how best to release open data requires a clear assessment of each dataset's value. This can be done by considering the potential uses and users of that data. Without such an assessment, it is impossible to weigh the privacy risks of a dataset (described in [Section 1.2](#)) against the potential value that the data creates.

Measuring the value of open data requires understanding how citizens use data. For datasets that have already been published, this can be done by analyzing the volume of usage and the value of those uses. To further learn how residents are using (or might use) data, cities can analyze public records requests, web traffic statistics, and online feedback forms. Perhaps more importantly, cities can also directly engage residents — including those who know little about open data — to learn more about how open data can provide value for the community.

When releasing new datasets, a broad understanding of open data uses is critical for predicting how that data might be used or even for tailoring the data to facilitate particular uses. Applications of open data from the past several years include:

- Social justice advocacy<sup>12</sup>
- Transit apps<sup>13</sup>
- Tools to explore a city's budget<sup>14</sup>
- Notifications to residents about events in their city<sup>15</sup>

---

<sup>12</sup>Becca James, "Stop and frisk in 4 cities: The importance of open police data," <https://sunlightfoundation.com/2015/03/02/stop-and-frisk-in-4-cities-the-importance-of-open-police-data-2/>.

<sup>13</sup>City and County of San Francisco, "Transportation << San Francisco Data," <http://apps.sfgov.org/showcase/apps-categories/transportation/>.

<sup>14</sup>City of Philadelphia, "Open Budget," <http://www.phila.gov/openbudget/>.

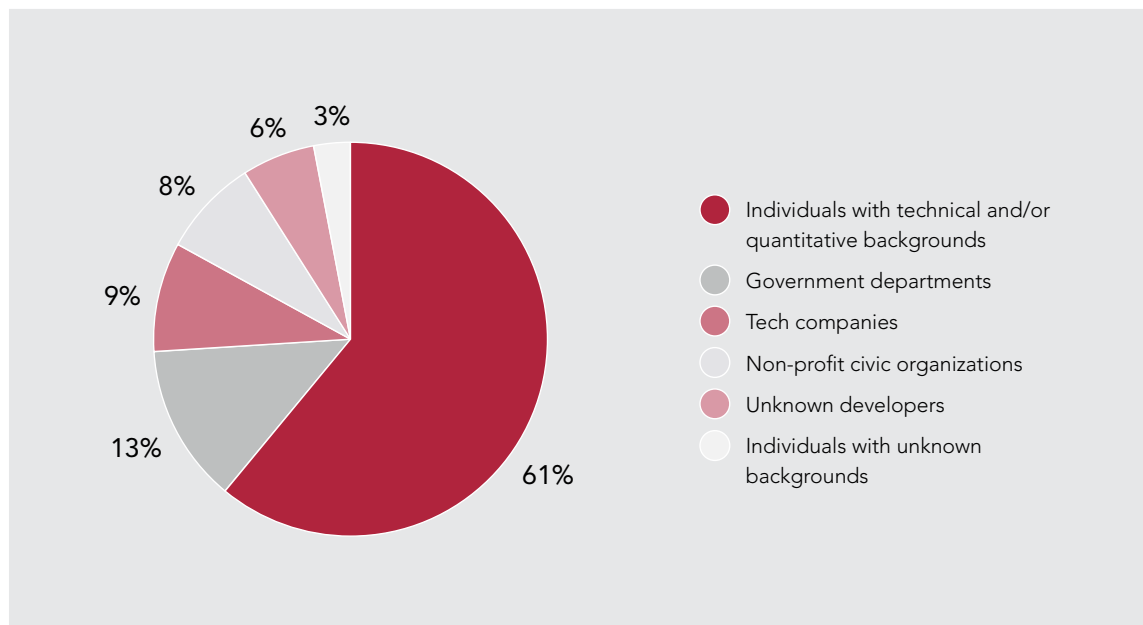
<sup>15</sup>"Citygram," <https://www.citygram.org>.

- Academic research about civic engagement<sup>16</sup>
- Visualizations of demographics in the US<sup>17</sup>

A 2016 study<sup>18</sup> of New York City's open data portal<sup>19</sup> surveyed 77 applications that utilized the City's data. Among these applications, the four most popular topics were 311 requests, crime, public transit, and the environment. To understand the benefits of open data, cities should also consider who is using open data; in addition to evaluating the uses and topics of the applications, this study also profiled the developers of these applications. The results, shown in *Figure 3*, indicate that most applications are built by "Individuals with technical and/or quantitative backgrounds" (such as civic hackers), but that government departments, tech companies, and nonprofits are also important developers.

The diversity of these applications and users suggests that there are many avenues for open data to provide value. To provide the most useful open data possible, and to properly weigh these benefits against potential privacy risks in the data, cities should assess and document every dataset before its release to determine how the data might be used, who will use it, and who will benefit from those uses.

*Figure 3. Background of developers using NYC open data*



Based on data from Karen Okamoto, "What is being done with open government data? An exploratory analysis of public uses of New York City open data," *Webology* 13, no. 1 (2016).

<sup>17</sup>Dustin Cable, "The Racial Dot Map: One Dot Per Person for the Entire United States," <http://demographics.coopercenter.org/DotMap/index.html>.

<sup>18</sup>Karen Okamoto, "What is being done with open government data? An exploratory analysis of public uses of New York City open data," *Webology* 13, no. 1 (2016).

<sup>19</sup>The City of New York, "NYC Open Data," <https://nycopendata.socrata.com>

# TAKE ACTION

The following form guides cities through benefit assessments. As shown, benefit is calculated at the asset level rather than the dataset level; this allows for tailored assessments that consider individual features that contribute to benefit and risk, and can be mitigated.

The following two sections (1.2 and 1.3) provide forms for risk and mitigation assessments. See the [Appendix](#) for a full risk-benefit analysis form that combines benefit, risk, and mitigation assessments.

DATA FEATURES (ASSETS)	ADVANTAGE EVENTS	ADVANTAGE SOURCES	BENEFIT																
<p><i>What are the rows, columns, entries, or sets of entries that may contribute to the overall benefit?</i></p>	<p><i>In what forms is the data feature beneficial? How will it be used?</i></p>	<p><i>Who might use the data feature?</i></p>	<p><i>What is the overall benefit of the data feature?</i></p>																
<p><b>Example 1:</b></p> <p>Pickup and dropoff locations for taxi trips</p>	<p><input checked="" type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregated data</p> <p>Potential uses:</p> <ul style="list-style-type: none"> <li>• Understand traffic patterns</li> <li>• Study working conditions of taxi drivers</li> </ul>	<p><input checked="" type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input checked="" type="checkbox"/> Individuals</p> <p><input checked="" type="checkbox"/> Journalists</p> <p><input checked="" type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p><b>LIKELIHOOD</b></p> <p>What is the probability that the impact will be realized?</p> <p><b>IMPACT</b></p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr> <td></td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>L</td> <td>L</td> <td>L</td> <td>M</td> </tr> <tr> <td>M</td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>H</td> <td>M</td> <td>H</td> <td>H</td> </tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H
	L	M	H																
L	L	L	M																
M	L	M	H																
H	M	H	H																
<p><b>Example 2:</b></p> <p>Sexual assault locations for 911 data</p>	<p><input checked="" type="checkbox"/> Individual records</p> <p><input checked="" type="checkbox"/> Aggregate data</p> <p>Potential uses:</p> <ul style="list-style-type: none"> <li>• Understand crime patterns</li> <li>• Look up details of specific cases</li> </ul>	<p><input checked="" type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input checked="" type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p><b>LIKELIHOOD</b></p> <p>What is the probability that the impact will be realized?</p> <p><b>IMPACT</b></p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr> <td></td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>L</td> <td>L</td> <td>L</td> <td>M</td> </tr> <tr> <td>M</td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>H</td> <td>M</td> <td>H</td> <td>H</td> </tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H
	L	M	H																
L	L	L	M																
M	L	M	H																
H	M	H	H																
	<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p><b>LIKELIHOOD</b></p> <p>What is the probability that the impact will be realized?</p> <p><b>IMPACT</b></p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr> <td></td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>L</td> <td>L</td> <td>L</td> <td>M</td> </tr> <tr> <td>M</td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>H</td> <td>M</td> <td>H</td> <td>H</td> </tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H
	L	M	H																
L	L	L	M																
M	L	M	H																
H	M	H	H																



One challenge for open data programs is prioritizing data for release: cities maintain many datasets that could be of public interest and value, and often struggle to determine the relative potential impact of these datasets. Given the work involved in releasing each dataset, it is important for cities to prioritize data effectively.

Mark Headd, the former Chief Data Officer for the City of Philadelphia, provides guidance for cities, writing, “governments should concentrate on The 3 B’s: Buses (transit data), Bullets (crime data) and Bucks (budget & expenditure data).”<sup>20</sup> Similarly, citing a desire to “focus on what matters,” Abhi Nemani, the former Chief Data Officer for the City of Los Angeles, adds four more categories to his list of the most essential data to release: bills, 211 (services), 311 (issues), and 411 (questions).<sup>21</sup> The strategy endorsed by Headd and Namani emphasizes an explicit focus on releasing data that will most likely lead to benefits.

Another challenge for open data programs is facilitating benefits of open data by helping the public realize the data’s potential uses. A recent study of open data and civic engagement in Cambridge, MA found that a “Crowdsourced Problem Inventory” was far and away the most heavily desired open data engagement tool; the report suggests that Cambridge should “create a ‘problem inventory’ that allows city staff and residents to scope out city needs, and share ideas and solutions.”<sup>22</sup> Such a strategy implies that open data portals should contain a repository of problems and questions that go along with each dataset. When a department shares data, it should also share relevant questions or analyses related to that data. Questions to analyze or requests (e.g., for a web application) based on the data could also come from community members. This approach would help the public translate data into valuable uses.

---

<sup>20</sup>Mark Headd, “In Defense of Transit Apps,” <https://civic.io/2014/06/13/in-defense-of-transit-apps/>.

<sup>21</sup>Abhi Nemani, “Small (City) Pieces, Loosely Joined,” <https://medium.com/@abhinemani/small-city-pieces-loosely-joined-5202fb5a93e3>.

<sup>22</sup>Jennifer Angarita and The City of Cambridge. “Amplifying Civic Innovation: Community Engagement Strategies for Open Data Collaborations.” (2016) [https://docs.google.com/viewerng/viewer?url=https://data.cambridgema.gov/api/file\\_data/f879b5f3-aa03-4e53-8600-7f5270299a62](https://docs.google.com/viewerng/viewer?url=https://data.cambridgema.gov/api/file_data/f879b5f3-aa03-4e53-8600-7f5270299a62).

## 1.2 RECOGNIZE THE LIMITS OF DE-IDENTIFICATION TECHNIQUES AND EVALUATE THE PRIVACY RISKS OF RELEASING DATA.

The privacy issues that can occur when open data is released may emerge due to a variety of data features. Cities should be mindful of vulnerabilities, threat sources, and threat events when evaluating the privacy risks involved in releasing open data.

The potential threat events from open data revealing private information are manifold. The following table describes several key consequences of sensitive information being released in open data.

THREAT EVENT	DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Re-identification</b>	Re-identification occurs when individual identities are inferred from data that has been de-identified (i.e., altered to remove individual identity from the data), and new information about those re-identified identities becomes known.	Re-identification involves the ability to learn information about individuals that would not otherwise be known. In many cases this new information can lead to a variety of harms for the re-identified individuals such as embarrassment, shame, identity theft, discrimination, and targeting for crime.	In 2000, Latanya Sweeney showed how de-identified health records could be combined with voting registration records to re-identify the health records of most individuals in the US. <sup>23</sup> This meant that it was possible to identify the individual referenced in many health records that were released under the assumption of anonymity. Scientific American describes a notable example: “William Weld, then the [Massachusetts] governor, assured the public that identifying individual patients in the records would be impossible. Within days, an envelope from a graduate student at the Massachusetts Institute of Technology arrived at Weld’s office. It contained the governor’s health records.” <sup>24</sup>
<b>False re-identification</b>	When data is partially anonymous, individuals are at risk of having sensitive facts incorrectly connected to them through flawed re-identification techniques. This is especially likely to occur when open data is of low quality, and contains incorrect information or is difficult to interpret.	Failed re-identification can be as troubling as successful re-identification. Individuals might have incorrect inferences made about them, which could lead to the same harms listed above for re-identification. These harms might be even more severe for false re-identification, since the outcomes will be based on false information or assumptions.	A release of data pertaining to 2013 taxi trips in New York City allowed journalists to determine where celebrities who had been photographed getting in or out of taxis were going to and coming from, along with the fare and tip paid. Surprisingly, many of these trips contained no recorded tip, leading to reports that certain celebrities were stingy and, in response, defenses from these celebrities’ agents. <sup>25</sup> Further analysis of the data revealed that many trips simply have no recorded tip, suggesting that the assumption that some celebrities paid no tip was in fact incorrect and due to issues with data quality.

<sup>23</sup>Latanya Sweeney. “Simple Demographics Often Identify People Uniquely.” (2000)

<sup>24</sup>Erica Klarreich, “Privacy by the Numbers: A New Approach to Safeguarding Data,” Quanta Magazine (2012).

<sup>25</sup>J.K. Trotter, “Public NYC Taxicab Database Lets You See How Celebrities Tip,” Gawker, October 23, 2014. <http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546>.

THREAT EVENT	DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Profile-building</b>	Many companies and other groups compile information about individuals to build a digital profile of each person's demographics, characteristics, habits, and preferences. Open data might contribute new information to these profiles.	Profiles built on data about individuals can be used to analyze and target information to specific segments of the population, thus facilitating algorithmic discrimination and exclusionary marketing.	It has become common practice for companies to target ads to users based on individual preferences, and, in some cases, treat customers differently based on profiles developed by compiling data about those individuals. Bloomberg calls this practice "Weblining, an Information Age version of that nasty old practice of redlining, where lenders and other businesses mark whole neighborhoods off-limits. Cyberspace doesn't have any real geography, but that's no impediment to Weblining. At its most benign, the practice could limit your choices in products or services, or force you to pay top dollar. In a more pernicious guise, Weblining may permanently close doors to you or your business." <sup>26</sup> Open data can contribute new information that feeds online profiles and allows for potential discrimination.
<b>Online discoverability</b>	Information that is available online and accessible from an online search.	When information in open data appears in online search results, it appears to a wide audience who might not otherwise have sought out that information. This is a significant change from the past, in which government records were typically available only to those who visited city hall to access them. Many citizens will be concerned when open data associated with their identity can be discovered through online searches for their name or address. Even if people are comfortable with the data being released on an open data portal, they might assume that the data is accessible only to those who seek it out. Exposing information in open data to online search engines can violate this assumption.	Multiple websites today post arrest records, including mug shots, to the Internet. <sup>27</sup> While this information is public record, traditionally one would have needed to go to a courthouse to obtain it. Now one can find this information, even inadvertently, just by searching the name of someone who is listed by mug shot websites. This is especially damaging, New York Times writes, because "Mug shots are merely artifacts of an arrest, not proof of a conviction, and many people whose images are now on display were never found guilty, or the charges against them were dropped. But these pictures can cause serious reputational damage." <sup>28</sup> The Times cites examples such as an individual who was denied a job due to online mug shots that appeared when a potential employer searched his name. These sites typically require fees up to several hundred dollars to have a mug shot removed, a practice that many have called extortion.

<sup>26</sup>Marcia Stepanek, "Weblining: Companies are using your personal data to limit your choices—and force you to pay more for products," Bloomberg April 3, 2000.

<sup>27</sup>"Mugshots," <http://mugshots.com/>

<sup>28</sup>David Segal, "Mugged by a Mug Shot Online," The New York Times, October 5, 2013. <http://www.nytimes.com/2013/10/06/business/mugged-by-a-mug-shot-online.html>.

THREAT EVENT	DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Public backlash</b>	Whenever sensitive information is published as open data, the public is likely to respond by blaming the government entity that released the data and losing faith in that entity to act as responsible data stewards.	Public disapproval of open data releases may result from one of the outcomes described above and suggest that the city is not acting with the best interests of its residents in mind. Furthermore, public disapproval detracts from the viability of an open data program. Without public trust in a city to responsibly share data, open data programs will struggle to gain necessary support for releasing data. More broadly, backlashes due to sensitive data releases undermine the public's trust in government.	In June 2016, Washington, DC published online the City's full voter list, which includes names, addresses, and political affiliations. <sup>29</sup> Many people responded with shock and outrage that DC would publish this information in such a widely available format, tweeting with hashtags like "#open_data_fail" <sup>30</sup> and calling the event "horrific." <sup>31</sup> While a public records law mandated that DC release this information, the event nonetheless made many individuals lose faith in DC as a responsible steward of their information. <sup>32</sup>

The negative outcomes described above can emerge through a variety of means: there are many ways that data can be mutated or combined to reveal sensitive information about individuals. Protecting privacy requires an understanding of how to identify and characterize sensitive information. The following table describes key vulnerabilities in municipal data that may allow individual privacy to be compromised.

VULNERABILITY	DATA DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Direct identifiers</b>	Features within a dataset that, on their own, identify individuals. These features (such as name, address, and Social Security Number) have traditionally been known as personally identifiable information (PII).	Because direct identifiers implicate an individual, all of the data tied to that identifier can be connected to the individual in question.	One dataset commonly released by open data programs is property assessments. Because this information includes each property's owner and address (direct identifiers), most records can be connected to an individual. Any information attached to these records (such as property value, renovation history, and violations) can therefore also be traced back to an individual.

<sup>29</sup>Ethan Chiel, "Why the D.C. government just publicly posted every D.C. voter's address online," Fusion, June 14, 2016. <http://fusion.net/story/314062/washington-dc-board-of-elections-publishes-addresses/>.

<sup>30</sup>Ashkan Soltan, <https://twitter.com/ashk4n/status/742466746079010817>.

<sup>31</sup>Jake Laperruque, <https://twitter.com/JakeLaperruque/status/742464398619512832>.

<sup>32</sup>Chiel, "Why the D.C. government just publicly posted every D.C. voter's address online."



VULNERABILITY	DATA DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Quasi (a.k.a. indirect) identifiers</b>	Features within a dataset that, in combination with other data, identify individuals. The ability to link features across datasets and learn about individuals is known as the mosaic effect.	Seemingly innocuous data can become revealing when combined with other datasets. Because quasi identifiers provide some information about individuals (although not enough by themselves to identify someone), they often facilitate linkage attacks (using the mosaic effect) that combine auxiliary information with quasi identifiers to identify individuals.	In a 2000 study, Latanya Sweeney showed how de-identified health records (containing the quasi identifiers birthdate, gender, and zip code about every individual) could be combined with voting registration records (which contain direct identifiers such as names along with the quasi identifiers mentioned above) to re-identify the health records of most individuals in the US. <sup>33</sup>
<b>Metadata (e.g., behavioral records)</b>	As The National Information Standards Organization describes, "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information." <sup>34</sup> In a database of emails, for example, metadata contains the sender, recipient, and timestamp of emails. While email metadata does not contain the contents of emails, it can reveal patterns about how people correspond. As such, metadata often comprises behavioral records.	While metadata has not traditionally been seen as sensitive, the President's Council of Advisors on Science and Technology (PCAST) writes, "There is no reason to believe that metadata raise fewer privacy concerns than the data they describe." <sup>35</sup> Although individual metadata records may appear anonymous, large sets of metadata describe detailed and unique patterns of behavior that make it possible to identify individuals and learn intimate details about those people. Behaviors of individuals can be discovered based on auxiliary knowledge (such as paparazzi photographs <sup>36</sup> ) or analyzing trends in the data (such as regular appearances at specific addresses <sup>37</sup> ). Furthermore, the privacy risks related to metadata are particularly troubling because such data can reveal intimate details of a person's life that would never otherwise be known and that the re-identified individual may never expect to be accessible.	Metadata is particularly sensitive when it is longitudinal, i.e., when multiple records of the same individual can be connected. In a 2015 study of de-identified credit card metadata, computer scientists showed that many people could be uniquely re-identified from records indicating the times and locations of each person's purchases. <sup>38</sup> Because people's movements and spending habits are idiosyncratic and unique, even a small number of records from one person are unlikely to be replicated by anyone else. In particular, the authors found that "knowing four random spatiotemporal points or tuples is enough to uniquely reidentify 90% of the individuals and to uncover all of their records." <sup>39</sup> Another study found that it was possible to predict people's personalities based on their mobile phone metadata. <sup>40</sup>
<b>Addresses</b>	Street addresses or location names.	Location data is often highly identifiable and can reveal particularly sensitive details about individuals. Because addresses identify where someone lives or where an event occurred, they are a rich source of information that make it easy to re-identify or learn intimate information about someone. Locations are also easy to link across datasets, facilitating the mosaic effect.	Many cities publish data about 311 requests, which relate to topics such as street and sidewalk repairs, missed trash pickups, animal waste, and pest complaints. Because a typical entry in a 311 dataset includes the address for which the request is made along with a description of the issue, many requests can be re-identified to determine the requester and information about that person's life.

<sup>33</sup>Sweeney, "Simple Demographics Often Identify People Uniquely."

<sup>34</sup>National Information Standards Organization. "Understanding Metadata." (2004) <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>.

<sup>35</sup>President's Council of Advisors on Science and Technology, "Big Data and Privacy: A Technological Perspective."

<sup>36</sup>Anthony Tockar, "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," <https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/>.

<sup>37</sup>Ibid.

<sup>38</sup>de Montjoye et al., "Unique in the shopping mall: On the reidentifiability of credit card metadata."

<sup>39</sup>Ibid.

<sup>40</sup>Yves-Alexandre de Montjoye et al., "Predicting personality using novel mobile phone-based metrics," in Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (Washington, DC: Springer, 2013).

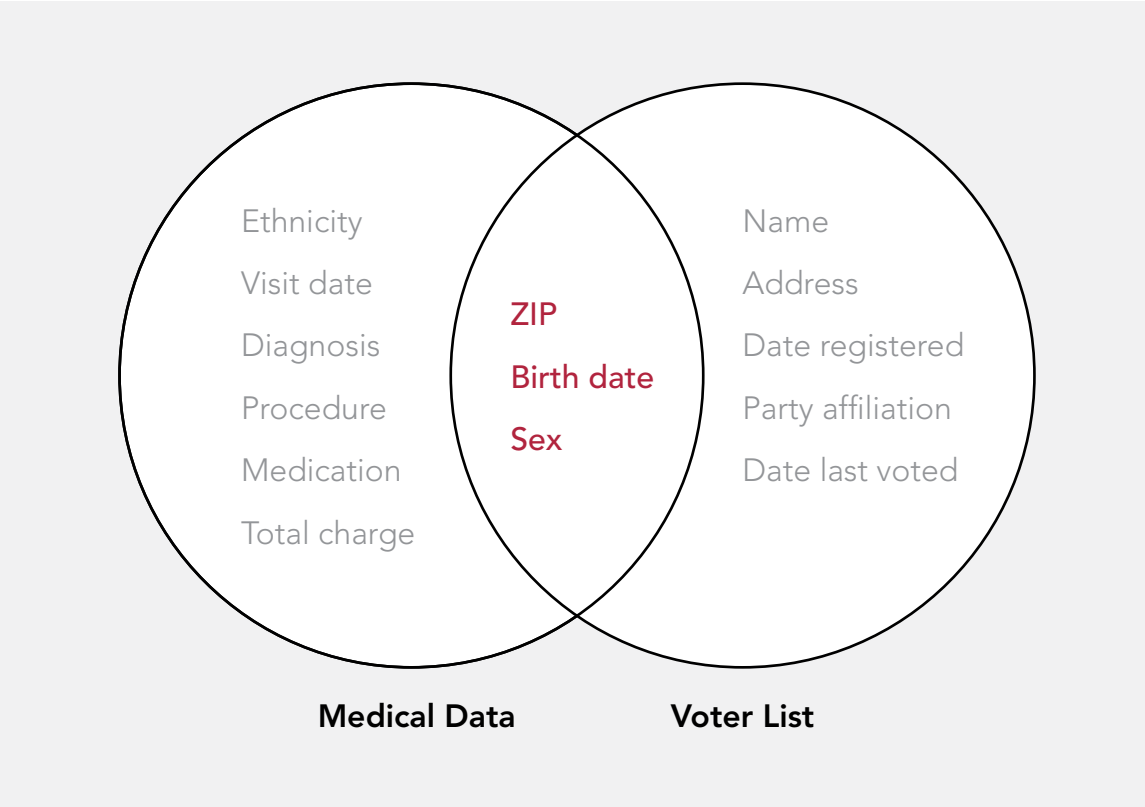
VULNERABILITY	DATA DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Geographic coordinates</b>	Coordinates that identify a unique location on a map (i.e., latitude and longitude).	Geographic coordinates present the same vulnerabilities as addresses since they translate into locations. Because geographic coordinates do not by themselves reveal a location, however, they may appear to be less sensitive than the addresses they represent. This is misleading, as it is simple to obtain an address from geographic coordinates through a process known as "reverse geocoding."	Crime data is one of the most heavily sought municipal datasets and, in the case of sexual assault-related incidents, one of the most sensitive. In order to protect the identities of victims when sharing open data, many jurisdictions remove the names and addresses associated with sexual assault incidents. However, such data occasionally includes the geographic coordinates of these incidents. Because it is relatively simple to obtain an address from geographic coordinates, this makes the victims of sexual assault highly identifiable. There are significant consequences if sexual assault victims are re-identified, including undue public scrutiny, violation of state shield laws, and potential chilling effects for future reports of sexual assault and domestic violence.
<b>Unstructured fields</b>	Fields that contain comments, descriptions, or other forms of unstructured text (as opposed to structured fields, in which entries must take one of several predetermined values). Photos can also be considered unstructured fields, as there are often few bounds on what information they may contain.	Freeform text fields are often used in unpredictable ways, meaning that their publication may expose unexpected sensitive information.	In 2012, Philadelphia's Department of Licenses & Inspections published gun permit appeals as part of its open data initiative. These permits included freeform text fields in which applicants explained why they needed the permit, and where some people wrote that they carry large sums of cash at night. <sup>41</sup> As a consequence for publishing 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." <sup>42</sup>
<b>Sensitive subsets</b>	Datasets can provide information about diverse populations or events. Each unique type of person or event represents a subset of the data.	Certain categories of people (such as minors and sexual assault victims) within a dataset may be systematically more sensitive than the rest. Information that might be suitable for release with the majority of data might be highly sensitive when it connects to these sensitive subsets.	In 2016, The Washington Post released a report describing how "the names of six people who complained of sexual assault were published online by Dallas police." <sup>43</sup> While the Dallas Police Department did not release "reports categorized as sexual assaults," some cases involving complaints of sexual assault were classified into categories such as "Class C Assault offenses" and "Injured Person." While it may be appropriate to release names in most cases in these general categories, the subsets related to sexual assault require special protections beyond what is needed for the majority of the data.

<sup>41</sup>Claudia Vargas, "City settles gun permit posting suit," The Philadelphia Inquirer, July 23, 2014.

<sup>42</sup>Ibid.

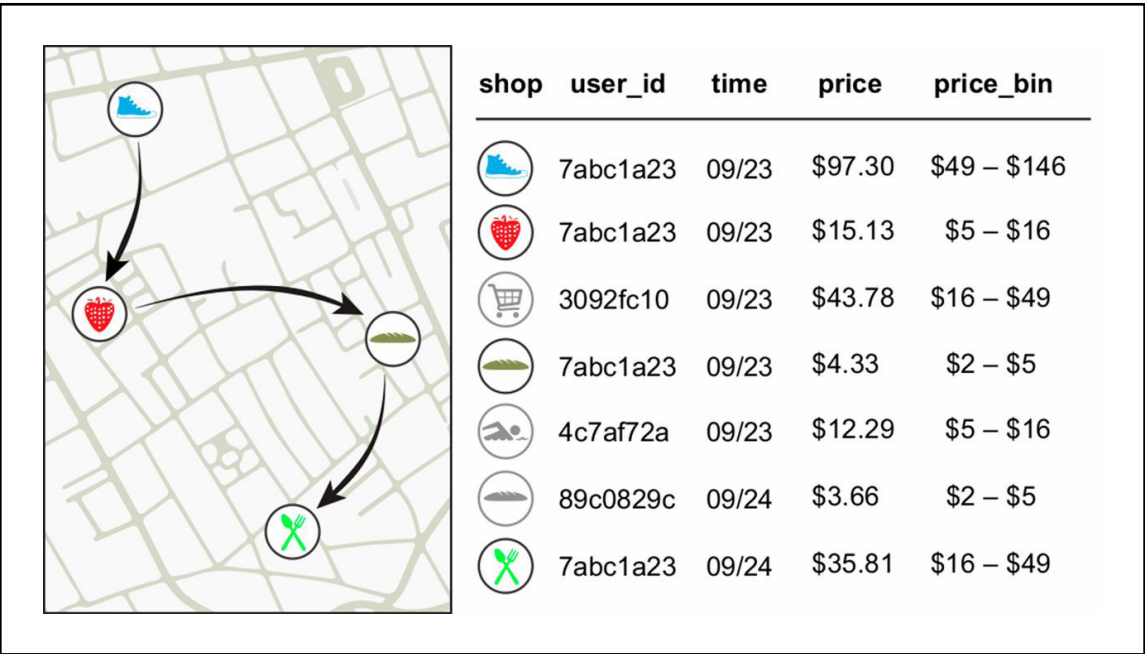
<sup>43</sup>Andrea Peterson, "Why the names of six people who complained of sexual assault were published online by Dallas police," The Washington Post, April 29, 2016. <https://www.washingtonpost.com/news/the-switch/wp/2016/04/29/why-the-names-of-six-people-who-complained-of-sexual-assault-were-published-online-by-dallas-police/>

Figure 4. Linking quasi-identifiers to re-identify data



Reproduced from Latanya Sweeney. "Simple Demographics Often Identify People Uniquely." (2000).

Figure 5. Financial metadata traces in a simply anonymized data set



From Yves-Alexandre de Montjoye et al., "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science* 347, no. 6221 (2015). Reprinted with permission from AAAS.

# TAKE ACTION

Before releasing data, cities must determine their risk tolerance and then evaluate the privacy risks present in the data. The following form is designed as a guide to the risk assessment process. This form should be completed in tandem with the benefit assessment form, so that every key data feature is evaluated for both benefit and risk. This process will simplify risk-benefit comparisons and facilitate the mitigation assessment ([Section 1.3](#)).

DATA FEATURES (VULNERABILITIES)	THREAT EVENTS	THREAT SOURCES	RISK																
<p><i>What are the rows, columns, entries, or sets of entries that may contribute to the overall risk?</i></p>	<p><i>In what ways is the data feature risky? How might it be abused?</i></p>	<p><i>Who might abuse the data feature?</i></p>	<p><i>What is the overall risk of the feature?</i></p>																
<p><b>Example 1:</b></p> <p>Pickup and dropoff locations for taxi trips</p>	<p><input checked="" type="checkbox"/> Individual records <input type="checkbox"/> Aggregated data</p> <p>Potential uses:</p> <ul style="list-style-type: none"> <li>• Re-identification of drivers</li> <li>• Re-identification of passengers</li> <li>• Unfair leverage for ride-share companies</li> </ul>	<p><input checked="" type="checkbox"/> Civic hackers <input type="checkbox"/> Community groups <input checked="" type="checkbox"/> Individuals <input type="checkbox"/> Journalists <input checked="" type="checkbox"/> Researchers <input type="checkbox"/> Other</p>	<p><b>LIKELIHOOD</b> What is the probability that the impact will be realized?</p> <p><b>IMPACT</b> What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr> <td></td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>L</td> <td>L</td> <td>L</td> <td>M</td> </tr> <tr> <td>M</td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>H</td> <td>M</td> <td>H</td> <td>H</td> </tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H
	L	M	H																
L	L	L	M																
M	L	M	H																
H	M	H	H																
<p><b>Example 2:</b></p> <p>Sexual assault locations for 911 data</p>	<p><input checked="" type="checkbox"/> Individual records <input checked="" type="checkbox"/> Aggregate data</p> <p>Potential uses:</p> <ul style="list-style-type: none"> <li>• Re-identification of victims</li> <li>• False re-identification of victims</li> </ul>	<p><input checked="" type="checkbox"/> Civic hackers <input checked="" type="checkbox"/> Community groups <input type="checkbox"/> Individuals <input checked="" type="checkbox"/> Journalists <input checked="" type="checkbox"/> Researchers <input type="checkbox"/> Other</p>	<p><b>LIKELIHOOD</b> What is the probability that the impact will be realized?</p> <p><b>IMPACT</b> What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr> <td></td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>L</td> <td>L</td> <td>L</td> <td>M</td> </tr> <tr> <td>M</td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>H</td> <td>M</td> <td>H</td> <td>H</td> </tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H
	L	M	H																
L	L	L	M																
M	L	M	H																
H	M	H	H																
	<p><input type="checkbox"/> Individual records <input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers <input type="checkbox"/> Community groups <input type="checkbox"/> Individuals <input type="checkbox"/> Journalists <input type="checkbox"/> Researchers <input type="checkbox"/> Other</p>	<p><b>LIKELIHOOD</b> What is the probability that the impact will be realized?</p> <p><b>IMPACT</b> What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr> <td></td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>L</td> <td>L</td> <td>L</td> <td>M</td> </tr> <tr> <td>M</td> <td>L</td> <td>M</td> <td>H</td> </tr> <tr> <td>H</td> <td>M</td> <td>H</td> <td>H</td> </tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H
	L	M	H																
L	L	L	M																
M	L	M	H																
H	M	H	H																



In August 2016, San Francisco released an Open Data Release Toolkit to guide thorough risk assessments of data being considered for publication.<sup>44</sup> This document is a great resource for evaluating the privacy risks of open data and determining how to publish municipal information.

The Open Data Release Form, shown below, outlines the Toolkit's process. Step 2 of the Release Form focuses on conducting a risk assessment of the data, and rests on assessing public expectations of privacy, the repercussions of re-identification, and likelihood of re-identification. The full Toolkit<sup>45</sup> provides further instructions about how to fill out the Release Form, as well as contextual information about open data privacy risks and how to mitigate them.

## Documenting the Decision-Making Process

The following Open Data Release Form should be used to document your decision-making.  
Refer to the remainder of this Toolkit for step-by-step guidance to fill out this form.

### Open Data Release Form

Basic Information																									
Department																									
Department contact																									
Contact details																									
Date																									
<b>Step 1: Identify sensitive or protected datasets</b>																									
1A. Dataset																									
1B. Relevant fields																									
<b>Step 2: Identifiability Risk Assessment</b>																									
<b>2A. Value of publication</b>	Value of publication <input type="checkbox"/> Low <input type="checkbox"/> Moderate <input type="checkbox"/> High																								
<b>2B-1. Risk of publication - impact</b>	<div>           (a) <b>Individual Expectation of Privacy</b>  <input type="checkbox"/> Low  <input type="checkbox"/> Moderate  <input type="checkbox"/> High         </div> <div>           (b) <b>Repercussions</b>  <input type="checkbox"/> No discernable  <input type="checkbox"/> Minor  <input type="checkbox"/> Moderate  <input type="checkbox"/> Major         </div> <div>           (c) <b>Impact = individual expectation of privacy X repercussions (legal, financial, etc.)</b> <table border="1"> <thead> <tr> <th rowspan="2">Impact Level</th> <th colspan="4">Repercussions</th> </tr> <tr> <th>No discernable</th> <th>Minor</th> <th>Moderate</th> <th>Major</th> </tr> </thead> <tbody> <tr> <td>Individual expectation of privacy - Low</td> <td>Very low</td> <td>Very low</td> <td>Low</td> <td>Moderate</td> </tr> <tr> <td>Moderate</td> <td>Very low</td> <td>Low</td> <td>Moderate</td> <td>Significant</td> </tr> <tr> <td>High</td> <td>Very low</td> <td>Moderate</td> <td>Significant</td> <td>High</td> </tr> </tbody> </table> <div> <input checked="" type="checkbox"/> Very low  <input type="checkbox"/> Low  <input type="checkbox"/> Moderate  <input type="checkbox"/> Significant  <input type="checkbox"/> High           </div> </div>	Impact Level	Repercussions				No discernable	Minor	Moderate	Major	Individual expectation of privacy - Low	Very low	Very low	Low	Moderate	Moderate	Very low	Low	Moderate	Significant	High	Very low	Moderate	Significant	High
Impact Level	Repercussions																								
	No discernable	Minor	Moderate	Major																					
Individual expectation of privacy - Low	Very low	Very low	Low	Moderate																					
Moderate	Very low	Low	Moderate	Significant																					
High	Very low	Moderate	Significant	High																					

DataSF.org: Open Data Release Toolkit - [Return to Top](#)

5 of 31



<sup>44</sup>Erica Finkle and DataSF. "Open Data Release Toolkit." (2016) <https://drive.google.com/file/d/0B0jc1tmJAITcR0RMV01PM2NyNDA/>.

<sup>45</sup>Ibid.

2B-2. Risk of publication - risk rating

(a) [Impact: See 2b-1\(c\) above](#)

(b) [Likelihood of re-identification attempt](#)

☐ Rare  
☐ Unlikely  
☐ Possible  
☐ Probable

(c) [Risk rating = impact X likelihood of re-identification attempt](#)

Risk Rating		Impact Level					
		Very low	Low	Moderate	Significant	High	Extreme
Likelihood	Rare	Very low	Very low	Low	Moderate	Significant	Significant
	Unlikely	Very low	Low	Moderate	Significant	High	High
	Possible	Very low	Moderate	Significant	High	Extreme	Extreme
	Probable	Very low	Significant	High	Extreme	Extreme	Extreme

☐ Very low  
☐ Low  
☐ Moderate  
☐ Significant  
☐ High  
☐ Extreme

2C. Weigh the value of publication against the risk of publication

Value v. Risk		Value		
		Low	Moderate	High
Risk Rating	Very low risk			
	Low risk			
	Moderate risk			
	Significant risk			
	High risk			
	Extreme risk			

☐ Moderate - high value. Very low - low risk  
☐ Low - high value. Very low - moderate risk  
☐ Low - high value. Low - significant risk  
☐ Low - high value. Moderate - high risk  
☐ Low - high value. Significant - extreme risk  
☐ Low - moderate value. High - extreme risk

**Step 3: Privacy Solutions**

3A. [Should the dataset be completely closed?](#)

☐ No  
☐ Yes  
 If "yes", do not proceed.

3B. [Identifiability spectrum level](#)

If the answer to Step 3A above is "no", then choose an identifiability spectrum level based on the results in Step 2C:

☐ Level 1: Readily identifiable data  
☐ Level 2: Masked data  
☐ Level 3: Obscured data  
☐ Level 4: Aggregate data

3C. [De-identification methods](#)

**Step 4: Accessibility Risk Assessment**

4A. [Assess likelihood of successful re-identification](#)

☐ Rare  
☐ Unlikely  
☐ Possible  
☐ Probable

4B. [Is the de-identified dataset still useful?](#)

☐ None  
☐ Low  
☐ Medium  
☐ High

4C. [Accessibility risk rating](#)

Risk Rating		Utility Level			
		High	Medium	Low	None
Likelihood	Rare	Very low	Very low	Low	Moderate
	Unlikely	Very low	Low	Moderate	Significant
	Possible	Low	Moderate	Significant	High
	Probable	Moderate	Significant	High	Extreme

☐ Very low  
☐ Low  
☐ Moderate  
☐ Significant  
☐ High  
☐ Extreme

4D. [Should the de-identified dataset be published?](#)

☐ Open  
☐ Limited Access  
☐ Closed

**Planning**

We plan to revisit the decisions in this form every...

☐ 6 months  
☐ 1 year  
☐ Other \_\_\_\_\_

Next date for review

**Notes**

# 1.3 CONSIDER A DIVERSITY OF POTENTIAL MITIGATIONS AND CHOOSE THE ONE BEST CALIBRATED TO THE SPECIFIC RISKS AND BENEFITS OF THE DATA.

Cities should develop a toolkit of mitigations — approaches for altering data to alleviate privacy risks. For each dataset, cities should select the mitigation best suited for the specific risks and benefits present.

Sensitive data often requires protection to make it appropriate as open data: when the risks of releasing data outweigh the benefits, cities may need to mitigate those risks before releasing data. Mitigations are controls that cities can use to protect sensitive information. Below is an overview of the most common techniques to consider when protecting individual privacy in open data. While no technique can remove all privacy risk involved in releasing data, the mitigations below outline important steps that cities can take to manage risks. In conjunction with the data-level approaches listed below, it is also possible to limit access to data through technical or contractual means such as access-control software or data-sharing agreements (see [Section 3.4](#)).

METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
Removing fields	Deleting fields that contain sensitive information.	Removing the addresses from every record in a dataset of police incidents.	Removing fields effectively removes the risks presented by those fields	This approach nullifies any utility made possible by the fields being removed. So the negative impact on utility is large when removing fields that enable valuable uses, but small for less valuable fields.
Removing records	Deleting records that are particularly sensitive, either because of the type of event represented or because of rare (and hence more easily identifiable) features.	Removing records of sexual assault from a dataset of police incidents.	This is an effective way to protect the privacy of those represented in the removed records.	Because only a subset of records have been removed and the rest remain intact, the data remains viable for analysis. However, the removal of records could skew the results or give a false impression about the underlying data. And any analyses that rely on the removed records will be negatively impacted.

METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
<b>Aggregating data</b>	Summarizing data across the population and releasing a report of those statistics.	Reporting the number of crimes that occurred each month rather than releasing data about individual incidents.	Releasing aggregated data effectively protects privacy, as no raw data entries are released.	This has a severe negative impact on utility, as there is no raw data allowing for insights beyond the statistics presented.
<b>Generalizing data</b>	Reducing the precision of fields in order to make each entry less unique.	Reporting addresses by hundred-block increments, block groups, or census tracts.	The less that data is generalized, the easier it is to re-identify someone. Lower levels of generalization (e.g., block group) provide more opportunities for re-identification than higher levels (e.g., zip code). However, while generalizing data can make re-identification more difficult, research has shown that coarsening data has only limited impact. <sup>46</sup>	The more that data is generalized and is characterized at less granular levels, the less useful it becomes. Lower levels of generalization (e.g., block group) provide more useful information than higher levels (e.g., zip code).
<b>k-anonymity<sup>47</sup></b>	Generalizing fields such that at least k individuals exhibit each feature within those fields. Different traits will require a different level of generalization, depending on how many other entries exhibit that trait.	For k=5, for example, generalizing dates of crime incidents such that every date shown contains at least five events that occurred. If 5 events occurred in a given hour, then the time of those events would be presented as the hour they occurred; if 5 events occurred in a given day, those events would be attributed to the day with no information about the time of day.	As with generalization, the improvement in privacy protection increases as the level of generalization (in this case, the value of k) increases. However, the efficacy of k-anonymity is limited in high-dimensional datasets (those that contain many fields) and in data that contains outliers or rare cases.	As with generalization, the negative impact on utility increases as the level of generalization (in this case, the value of k) increases

<sup>46</sup>Yves-Alexandre de Montjoye et al., "Unique in the Crowd: The privacy bounds of human mobility," Scientific Reports 3 (2013); de Montjoye et al., "Unique in the shopping mall: On the reidentifiability of credit card metadata."

<sup>47</sup>Latanya Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, no. 5 (2002).

METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
<b>Adding noise (a.k.a. random perturbation)<sup>48</sup></b>	Adjusting data with randomness to offset its original information.	Offsetting geographic coordinates of crime locations by a random distance and direction (generated from probability distributions).	The level of privacy protection increases as more noise is added to a dataset. The impact of noise depends on the density of the population in question: less dense populations require more noise to protect privacy.	As more noise is added to a dataset, the less useful it becomes, since the data presented becomes further removed from the events they represent. Furthermore, noisy data can be hard to communicate to the public and may be seen as misleading or an obfuscation of the truth. <i>We recommend against adding noise, and instead suggest generalizing data.</i>
<b>Creating anonymous identifiers</b>	Replacing attributes with randomly generated codes that have no underlying connection to the attribute they replace. This is done through a correspondence table, in which each unique attribute is paired with a random identifier that will replace that attribute wherever it appears.	In a dataset of taxi trips, replacing each unique license plate with its own unique ID number (e.g., a random number drawn from between 1 and the total number of license plates). Every entry containing a given license plate would have the same ID number.	Anonymous IDs can help protect privacy, assuming that the anonymous IDs are randomly generated and have no systematic connection to the attributes they replace (which would occur for example if the numbers were assigned based on the alphanumeric order of license plates or a direct hash of license plates). Note that creating anonymous IDs does not protect against re-identifications or inferences based on analyzing patterns of behavior. Furthermore, having any common identifier across all entries related to a specific individual means that once one entry has been re-identified, all entries for that person have also been re-identified.	This approach should have minimal impacts on utility, since it is still possible to track attributes across records.

<sup>48</sup>Dave Stinchcomb, "Procedures for Geomasking to Protect Patient Confidentiality," in ESRI International Health GIS Conference (Washington, D.C.2004).

METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
<b>Differential privacy<sup>49</sup></b>	<p>Differential privacy is a formal mathematical definition of privacy that provides a provable guarantee of privacy against a wide range of potential attacks.<sup>50, 51</sup> It is not a single tool, but rather a standard of privacy that 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 to be released and used).</p> <p>Theoretical research on differential privacy is rapidly advancing, and the number of practical tools providing differential privacy is continually growing. For these reasons, differentially private tools are becoming an increasingly promising solution for cities to use in combination with other legal and technological tools for sharing data while protecting individual privacy.</p>	<p>Government agencies and corporations currently use differentially private tools to provide strong privacy protection when sharing statistics.<sup>52, 53</sup> The Census Bureau, for example, currently makes some of its data available using non-interactive differentially privacy.<sup>54</sup> Additional tools for differentially private analysis are under development at research institutions.<sup>55, 56</sup></p>	<p>Differential privacy provides strong guarantees regarding the exact level of privacy risk available through a dataset. In contrast to traditional de-identification techniques that are often designed to address a narrow class of attacks, systems that adhere to strong formal standards like differential privacy provide protection that is robust to a wide range of potential attacks — including attacks that are unknown at the time of deployment — and do not require the person applying the technique to anticipate particular modes of attack.</p>	<p>Differential privacy minimally alters the underlying data, ensuring that the data retains almost all of its utility even after transformation. This feature distinguishes differentially private tools from traditional de-identification techniques, which often require more blunt alterations.</p> <p>In addition to their robust privacy guarantee, differentially private tools have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. Nonetheless, as with the approach above of adding noise, users of differentially private results may struggle to interpret the data. Furthermore, providing data transformed in this way limits the ability to review specific records and might be seen as antithetical to open data.</p>

<sup>49</sup>Alexandra Wood and Micah Altman, Personal communication, 2016.

<sup>50</sup>Cynthia Dwork, "A firm foundation for private data analysis," Communications of the ACM 54, no. 1 (2011).

<sup>51</sup>Klarreich, "Privacy by the Numbers: A New Approach to Safeguarding Data."

<sup>52</sup>U.S. Census Bureau Center for Economic Studies, "OnTheMap," <http://onthemap.ces.census.gov>.

<sup>53</sup>Andrew Eland, "Tackling Urban Mobility with Technology," Google Europe Blog, November 18, 2015. <https://europe.googleblog.com/2015/11/tackling-urban-mobility-with-technology.html>.

<sup>54</sup>U.S. Census Bureau Center for Economic Studies, "OnTheMap".

<sup>55</sup>Privacy Tools for Sharing Research Data, "Private Data Sharing Interface," <https://beta.dataverse.org/custom/DifferentialPrivacyPrototype/>.

<sup>56</sup>University of Pennsylvania, "Putting Differential Privacy to Work," <http://privacy.cis.upenn.edu/index.html>.



# TAKE ACTION

If the results of the benefit and risk assessments imply that the data should not be released in its present state, cities must mitigate those privacy risks before releasing data. The following form is designed to guide cities through selecting and evaluating potential mitigations.

DATA FEATURES (ASSETS AND VULNERABILITIES)  What are the rows, columns, entries, or sets of entries that may contribute to benefit or risk?	RISK-BENEFIT RATIO  What is the overall risk-benefit ratio as determined by the benefit assessment (Section 1.1) and the risk assessment (Section 1.2)?	MITIGATIONS  What are the potential controls to mitigate risk?	RISK-BENEFIT RATIO AFTER MITIGATION  What is the outcome of the risk-benefit analysis after mitigation?	FINAL OUTCOME  What is the final decision for how to release the data?																																
<b>Example 1:</b>  Pickup and dropoff locations for taxi trips	<div>BENEFIT</div> <table><tr><td></td><td>L</td><td>M</td><td>H</td></tr><tr><td>L</td><td>M</td><td>L</td><td>L</td></tr><tr><td>M</td><td>H</td><td>M</td><td>L</td></tr><tr><td>H</td><td>H</td><td>H</td><td>M</td></tr></table> <div>RISK</div>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M	<div><input checked="" type="checkbox"/> Remove fields</div> <div><input type="checkbox"/> Remove records</div> <div><input type="checkbox"/> Aggregate data</div> <div><input checked="" type="checkbox"/> Generalize data</div> <div><input type="checkbox"/> Anonymize IDs</div> <div><input type="checkbox"/> Other</div> <div>Mitigation chosen:</div> <div>Generalize locations to the block group</div>	<div>BENEFIT</div> <table><tr><td></td><td>L</td><td>M</td><td>H</td></tr><tr><td>L</td><td>M</td><td>L</td><td>L</td></tr><tr><td>M</td><td>H</td><td>M</td><td>L</td></tr><tr><td>H</td><td>H</td><td>H</td><td>M</td></tr></table> <div>RISK</div>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M	Release taxi data with pickup and dropoff locations generalized to the block group.
	L	M	H																																	
L	M	L	L																																	
M	H	M	L																																	
H	H	H	M																																	
	L	M	H																																	
L	M	L	L																																	
M	H	M	L																																	
H	H	H	M																																	
<b>Example 2:</b>  Sexual assault locations for 911 data	<div>BENEFIT</div> <table><tr><td></td><td>L</td><td>M</td><td>H</td></tr><tr><td>L</td><td>M</td><td>L</td><td>L</td></tr><tr><td>M</td><td>H</td><td>M</td><td>L</td></tr><tr><td>H</td><td>H</td><td>H</td><td>M</td></tr></table> <div>RISK</div>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M	<div><input checked="" type="checkbox"/> Remove fields</div> <div><input checked="" type="checkbox"/> Remove records</div> <div><input type="checkbox"/> Aggregate data</div> <div><input checked="" type="checkbox"/> Generalize data</div> <div><input type="checkbox"/> Anonymize IDs</div> <div><input type="checkbox"/> Other</div> <div>Mitigation chosen:</div> <div>Remove location fields</div>	<div>BENEFIT</div> <table><tr><td></td><td>L</td><td>M</td><td>H</td></tr><tr><td>L</td><td>M</td><td>L</td><td>L</td></tr><tr><td>M</td><td>H</td><td>M</td><td>L</td></tr><tr><td>H</td><td>H</td><td>H</td><td>M</td></tr></table> <div>RISK</div>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M	Release crime reports with the locations of sexual assault incidents removed.
	L	M	H																																	
L	M	L	L																																	
M	H	M	L																																	
H	H	H	M																																	
	L	M	H																																	
L	M	L	L																																	
M	H	M	L																																	
H	H	H	M																																	
	<div>BENEFIT</div> <table><tr><td></td><td>L</td><td>M</td><td>H</td></tr><tr><td>L</td><td>M</td><td>L</td><td>L</td></tr><tr><td>M</td><td>H</td><td>M</td><td>L</td></tr><tr><td>H</td><td>H</td><td>H</td><td>M</td></tr></table> <div>RISK</div>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M	<div><input type="checkbox"/> Remove fields</div> <div><input type="checkbox"/> Remove records</div> <div><input type="checkbox"/> Aggregate data</div> <div><input type="checkbox"/> Generalize data</div> <div><input type="checkbox"/> Anonymize IDs</div> <div><input type="checkbox"/> Other</div> <div>Mitigation chosen:</div>	<div>BENEFIT</div> <table><tr><td></td><td>L</td><td>M</td><td>H</td></tr><tr><td>L</td><td>M</td><td>L</td><td>L</td></tr><tr><td>M</td><td>H</td><td>M</td><td>L</td></tr><tr><td>H</td><td>H</td><td>H</td><td>M</td></tr></table> <div>RISK</div>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M	
	L	M	H																																	
L	M	L	L																																	
M	H	M	L																																	
H	H	H	M																																	
	L	M	H																																	
L	M	L	L																																	
M	H	M	L																																	
H	H	H	M																																	

# IN PRACTICE

In 2014, in response to a Freedom of Information Law (FOIL) request, New York City released data detailing every taxi ride recorded in registered NYC taxis during 2013.<sup>57</sup> The data contained information about pickup time and location, drop-off time and location, and the taxicab (in the form of license plate) and driver (in the form of medallion number) involved in each trip. In order to protect the identity of taxi drivers, NYC provided codes corresponding to the medallion and license numbers rather than the raw numbers.

When analyzing this information, data scientist Vijay Pandurangan realized that the codes for medallion and license plate represented hashes of the medallion and license numbers, generated using the hash function MD5 (an encryption function designed to turn identifying inputs into anonymized outputs in a manner that cannot be reverse engineered).<sup>58</sup> Knowledge of how the codes were generated, combined with the standard form of license plates and medallion numbers, compromised this protection, however. Pandurangan writes:

“A cryptographically secure hashing function, like MD5 is a one-way function: it always turns the same input to the same output, but given the output, it’s pretty hard to figure out what the input was as long as you don’t know anything about what the input might look like. This is mostly what you’d like out of an anonymization function. The problem, however, is that in this case we know a lot about what the inputs look like.”<sup>59</sup>

With this knowledge it was relatively simple for Pandurangan to re-identify the entire dataset in a matter of hours, thus revealing the identity, annual income, and home address of individual drivers.

This story shows that it is not sufficient merely to identify a mitigation approach that protects privacy. Instead, care must be taken to ensure that the chosen mitigation technique is tailored to the privacy risks present in a particular dataset, and should be tested to ensure robustness against potential re-identification attacks. In this case, it would have been more effective to de-identify the license and medallion numbers by replacing each with a random number (i.e., anonymizing the IDs). Because these new numbers would be assigned to the license and medallion numbers at random, without any systematic connection to the structure or order of the original data, it would have been impossible to reverse engineer the true numbers from the random ones. Alternatively, NYC could have released the data without including license or medallion numbers at all.

---

<sup>57</sup>Chris Whong, “FOILing NYC’s Taxi Trip Data,” [http://chriswhong.com/open-data/foil\\_nyc\\_taxi/](http://chriswhong.com/open-data/foil_nyc_taxi/).

<sup>58</sup>Vijay Pandurangan, “On Taxis and Rainbows,” <https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1>

<sup>59</sup>Ibid.

## 2. CONSIDER PRIVACY AT EACH STAGE OF THE DATA LIFECYCLE.

Every stage of the data lifecycle involves actions and decisions that have consequences for individual privacy. Responsibly opening data to the public involves a process far more complex than just uploading data.

Recognizing the limits of de-identification (as described in [Section 1.2](#)), cities should focus on mitigating privacy risks throughout the data's entire lifetime rather than only when releasing data. Privacy should not be an isolated issue, considered only when data is about to be released. In this vein, Ira Rubinstein and Woodrow Hartzog argue that privacy laws should be "process-based, contextual, and tolerant of harm, so long as procedures to minimize risk are implemented ex ante."<sup>60</sup> Rather than focus only on outcomes, effective privacy protection requires a thorough process that minimizes risks before issues arise.

Data, like any other type of infrastructure, requires effective management at all stages of its lifecycle. Given that open data involves releasing data, the preparation of data for public consumption has traditionally been the focus of privacy protection efforts related to open data. While this process is important (and is discussed in [Section 2.3](#)), effective privacy management matters at all stages of the data lifecycle.

The lifecycle of open data can be summarized into the following stages:

1. Collect data
2. Maintain data
3. Release data
4. Delete data

These stages all have important implications for data privacy. This chapter describes the steps that should be taken within each stage to mitigate the privacy risks of open data.

---

<sup>60</sup>Ira S Rubinstein and Woodrow Hartzog, "Anonymization and Risk," *Washington Law Review* 91 (2016).

## 2.1 COLLECT: BE MINDFUL OF PRIVACY BEFORE COLLECTING ANY DATA.

Once data has been collected, it is susceptible to public release either as open data or through responses to public records requests. Limiting collection is often the best way to limit future disclosure.

Privacy risks are first created when data is collected: individuals cannot be re-identified unless data about them has been collected and stored in the first place. Furthermore, the more data that is collected the greater the likelihood and consequences of re-identification. This is true not just for the number of features measured but also the number of records collected; for longitudinal data such as taxi trips, every additional trip tracked adds to the risks present in the data.

Once data has been collected, it is vulnerable to public release through multiple means, at which point it can be used for re-identification. This risk is made especially acute due to public records laws, which can compel the disclosure of data that might otherwise not be proactively released as open data. Cities should therefore assume that any data they collect could be made public and should thoughtfully ensure that the benefits of collection outweigh the risks before data is gathered.

Finally, data collection by itself can be seen as a violation of individual privacy rights, even if the data is never made public. Many people are concerned about government surveillance – especially in recent years, due to the Edward Snowden revelations – and are therefore likely to be bothered by anything seen as excessive government data collection, particularly in the absence of any communication about the benefits of such collection (for more on public perceptions of data privacy, see [Chapter 4](#)).

## TAKE ACTION

Many privacy risks arise due to excessive collection of features or records that are not essential for utilizing the data but provide information that can be used to identify an individual or one's behavior. Cities should therefore follow the guidelines of data minimization: "the practice of limiting the collection of personal information to that which is directly relevant and necessary to accomplish a specified purpose."<sup>61</sup> This practice limits the privacy risks involved in collecting data and makes it more likely that benefits of collection outweigh the risks.

The following exercise can help cities map data collection to impacts – both positive and negative – in order to make informed collection choices. By answering these questions, cities can evaluate the risks and benefits of collecting each data element and determine their data collection strategy.

To provide a sense of how this chart could enable informed data collection decisions, two rows have been filled out using potential features that could be collected as part of a public Wi-Fi program.

DATA FEATURE	BENEFITS OF COLLECTION	RE-IDENTIFICATION RISK	PUBLIC RESPONSE RISK	WOULD BE PUBLIC RECORD?	COLLECT?
	<i>What are the tangible benefits for which this data feature is crucial?</i>	<i>What information could be learned about individuals using this information?</i>	<i>Would collecting this information violate public expectations about acceptable government data collection? To what extent?</i>	<i>Would the data be public record? (Public records status compounds privacy risks, since it means that the data is more likely to be made public.)</i>	
<b>Domains visited (aggregated over all users)</b>	Allows operators to monitor service and understand how users take advantage the service.	Low. It would be difficult to identify anyone if the domains visited are not linked with the user.	Low. This is a justified collection to operate the Wi-Fi network and is disaggregated from any individual.	Yes.	Yes. This data will be useful and does not contain sensitive information.
<b>MAC addresses of passers-by</b>	Would allow analyses of how people move through the city	High. Using these data, it would be possible to track people's movements throughout the city.	High. This data would be collected from individuals who did not agree to any terms and are not using the service provided.	Probably, depending on jurisdiction and interpretation of open records laws.	No. This data could be useful, but involves collecting sensitive information and likely generating public pushback.

<sup>61</sup>Bernard Marr, "Why Data Minimization Is An Important Concept In The Age Of Big Data," *Forbes*, March 16, 2016. <http://www.forbes.com/sites/bernardmarr/2016/03/16/why-data-minimization-is-an-important-concept-in-the-age-of-big-data/>

# IN PRACTICE

Among of the key types of data collected by the sensors in Chicago's Array of Things project are pedestrian, bicycle, and automobile counts, which are calculated by analyzing with computer vision algorithms images taken by a camera on every sensor. While pedestrian and vehicle counts are not themselves sensitive, the images necessary to compute these metrics are: because the images can include identifiers such as faces and license plates, an extensive backlog of pictures or video from all the City's sensors would pose a significant privacy risk.

Recognizing these risks, the City of Chicago developed a plan to minimize the collection of these images by computing traffic counts on-board the sensors themselves, and then discarding the images and storing only the counts. This means that the images never need to be stored outside the cameras or beyond the time it takes to calculate the desired metrics.

In addition, while it is necessary to collect and store some images for the purpose of calibrating the sensors, the amount of content is limited to only the quantity necessary for this purpose, and access to the images is severely restricted. The program's privacy policy<sup>62</sup> describes its approach as follows:

"For the purposes of instrument calibration, testing, and software enhancement, images and audio files that may contain non-sensitive PII will be periodically collected to improve, develop, and enhance algorithms that could detect and report on conditions such as noted above. This raw calibration data will be stored in a secure facility for processing only by authorized researchers during the course of the Array of Things project, including for purposes of improving the technology to protect this non-sensitive PII. Access to this limited volume of data is restricted to operator employees, contractors and approved scientific partners who need to process the data for instrument design and calibration purposes. All individuals with access to this data will be subject to strict contractual confidentiality obligations and will be subject to discipline and/or termination if they fail to meet these obligations."

The Array of Things' approach to managing these images responsibly restricts privacy risks by minimizing data collection while simultaneously retaining all benefits from having the cameras in place.

---

<sup>62</sup>Array of Things, "Array of Things Operating Policies," <https://arrayofthings.github.io/final-policies.html>.



## 2.2 MAINTAIN: KEEP TRACK OF PRIVACY RISKS IN ALL DATA STORED AND MAINTAINED.

Managing privacy requires knowing what data the city possesses and what privacy risks that data poses. Data inventories and privacy grading schemas are effective ways to enable the auditing of data for privacy risks.

Among the biggest challenges in protecting privacy is keeping track of data and privacy risks. Cities maintain numerous datasets — often distributed across many different departments — making it difficult for cities to keep track of their many data resources. Without a comprehensive knowledge of available datasets, cities may make poor data management decisions or undertake redundant data collection efforts.

Furthermore, unknown and insufficiently monitored datasets pose privacy risks: it is impossible to mitigate risks that no one knows exists. Sensitive data that is not actively monitored will not be properly protected, increasing the likelihood that private information will be disclosed as open data or through public records requests. Personnel turnover and regular upgrades to records management systems add to the difficulty of properly evaluating an old dataset for privacy risks. The rapid pace of developments in data analytics is especially troubling, as it means that the risks in a dataset are constantly evolving even when the data remains unchanged.

# TAKE ACTION

It is critical for cities to maintain a detailed inventory of existing data in order to evaluate their data ecosystem and streamline privacy protection at each phase of the data lifecycle.

Data inventories should ensure that datasets contain thorough metadata, answering questions such as:

- What do the fields and entries mean?
- What internal process does the data represent? How is the data generated?
- What records management system generates the data?
- When, how, and why was this dataset created?
- How does this dataset get used?
- What other datasets contain similar, redundant, or complementary information?
- How frequently is the dataset updated?
- What department(s) and individual(s) are responsible for the data?
- How long is the mandated retention period? How long does the city intend to retain the data?

This documented information is necessary to enable cities to fully investigate their open data ecosystem. Such an inventory will also enable more informed and proactive uses of data throughout city hall, and will be massively helpful to consumers of any datasets that are published as open data.

The other key component for a data inventory is evaluating the privacy risks within each dataset. One potential approach for incorporating privacy into inventories is to develop grading schemas that classify the privacy risks included in every dataset into a small number of categories, such as High, Medium, and Low Risk. Such a system could provide a bird’s-eye view of privacy risks, helping the city target its resources to mitigate risks and identify good candidate datasets for release as open data.

		Damage (to individuals)		
		LOW: Minimally sensitive information is contained or could be revealed	MEDIUM: Mildly sensitive information is contained or could be revealed	HIGH: Very sensitive information is contained or could be revealed
Scale (of inference)	LOW: Targeted effort required for each individual, requires the use of unique data	Low risk		
	MEDIUM: A single method can be used to learn about some people in the data, and uses data that is available for some people		Medium risk	
	HIGH: A single method can be used to learn about most people in the data, and uses data that is available for most people			High risk

Notwithstanding the appeal of such a grading system, cities should be aware of several pitfalls inherent in developing a privacy grading system. In particular, the risk scores may not:

- Provide a full picture of risks and mitigations. It might be simple to make some high-risk datasets suitable for release, but difficult for others.
- Account for other data that exists and the mosaic effect. Even if a dataset on its own is not sensitive, it could pose risks when combined with other open datasets.
- Consider constituents' expectations and the relationship between the public and the city. A dataset considered high risk by the community in one city might be considered medium risk in another (depending on factors such as previous events related to data, the local tech community, and the social standing of the government).

With regard to this final point, developing an open data privacy schema presents a perfect opportunity to engage the public with respect to priorities related to privacy and open data. Asking the public for input is a great way to educate citizens about these issues, gauge their priorities, and collaboratively develop standards (see [Chapter 4](#)).

# IN PRACTICE

Some cities, states, and the federal government have already developed definitions that are used to classify datasets. A few such schemas are summarized below.

	CLOSED	INTERMEDIATE	OPEN
<b>San Francisco</b> <sup>63</sup>	Protected: this data is protected by law or regulation and can only be shared or accessed internally and per organizational procedures; OR this information includes individually identified information.	Sensitive: in its raw form, this data poses security concerns, could be misused to target individuals or poses other concerns.	Public: this data could be publicly disseminated without any concerns
<b>Washington State</b> <sup>64</sup>	Confidential Information: data is specifically protected from disclosure by law.  Confidential Information Requiring Special Handling: data is specifically protected from disclosure by law and subject to strict handling requirements dictated by statutes, regulations, or legal agreements.	Sensitive Information: data may not be specifically protected from disclosure by law and is for official use only. Sensitive information is generally not released to the public unless specifically requested.	Public Information: data can be or currently is released to the public. It does not need protection from unauthorized disclosure, but does need integrity and availability protection controls.
<b>Data.gov</b> <sup>65</sup>	Non-public: data could never be made available to the public for privacy, security, or other reasons as determined by your agency.	Restricted Public: data is only available under certain conditions or to certain audiences (such as researchers who sign a waiver)	Public: data are or could be made publicly available to all without restrictions

A related data classification scheme is Datatags,<sup>66</sup> a system developed by Latanya Sweeney at Harvard University to help researchers share data without violating privacy laws. Datatags provides users with an interactive survey that asks questions about a dataset to determine its level of sensitivity. Data is then classified into one of six categories, ranging from “Non-confidential information” to “Maximum sensitive personal information.”<sup>67</sup> By guiding users through the process of analyzing a dataset for privacy concerns, Datatags provides a simple interface to help those without expertise in data privacy manage how they share data. Because Datatags is designed to help manage data with respect to a specific set of privacy laws, it does not cover privacy risks that may fall outside the coverage of existing legislation. Nonetheless, it provides a valuable model for how to develop a simple, user-friendly tool for navigating complex data privacy decisions.

<sup>64</sup>Washington State Office of the Chief Information Officer. “Securing Information Technology Assets.” (2013)

<sup>65</sup>Project Open Data, “Project Open Data Metadata Schema v1.1,” <https://project-open-data.cio.gov/v1.1/schema/#accessLevel>

<sup>66</sup>Privacy Tools for Sharing Research Data, “DataTags,” <http://datatags.org>.

<sup>67</sup>“DataTags-Compliant Repositories,” <http://datatags.org/datatags-compliant>.

## 2.3 RELEASE: EVALUATE DATASETS FOR PRIVACY RISKS AND MITIGATE THOSE RISKS BEFORE RELEASING DATA.

Open data can pose privacy risks in a variety of ways. Responsibly releasing data requires a thorough analysis of privacy risks and the use of a wide range of de-identification techniques.

Preparing datasets to be released is one of the most common challenges faced by open data initiatives, as it is not always clear if the data poses privacy risks and, if so, how to manage those risks. As described in [Section 1.2](#), open data can compromise individual privacy in numerous ways that involve a variety of data types. It is only possible to mitigate these risks by first identifying the risks that are present. This section provides guidelines for evaluating datasets for privacy risks and sensitive information, and what de-identification techniques may be effective at mitigating these risks.

# TAKE ACTION

The following questionnaire is designed to help data owners identify and mitigate the privacy risks of a dataset before sharing it. It aims to highlight the diverse privacy risks that can exist within the datasets that cities are likely to share. Because of the highly contextual nature of open data privacy risks, the mitigating actions listed are suggestive rather than prescriptive; any determination of exactly how to prepare datasets for release must be made based on the particular risks and benefits involved.

	ATTRIBUTE	RISK DESCRIPTION	EXAMPLE(S)	MITIGATING ACTION(S)
<b>Category 1: Individual identifiers</b>	Does the data contain information and attributes directly tied to an individual?	Many types of information can be used to identify individuals within a dataset. Even if a field does not by itself identify an individual, it can be used in conjunction with other fields to do so.	<ul style="list-style-type: none"> <li>• Name</li> <li>• Sex</li> <li>• Race</li> <li>• Address</li> <li>• Birthdate</li> <li>• Phone number</li> <li>• User ID</li> <li>• License plate</li> </ul>	Reduce the precision of these fields or remove them entirely.
	Does the data contain repeated records of an individual's actions?	Behavioral records, often known as metadata, describe detailed and unique patterns of behavior that make it easy to identify individuals and learn intimate details about that person.	<ul style="list-style-type: none"> <li>• User IDs in records of bikeshare usage</li> <li>• License plates in records of taxi trips</li> </ul>	Remove the fields that provide references to individuals, so that records cannot be connected based on the person. Or provide anonymous identifiers in place of these individual IDs, ensuring that they are randomly generated and there is no systematic connection between the original and anonymized IDs (such as alphabetical order).
<b>Category 2: References to location</b>	Does the dataset contain references to locations?	Location data is often highly identifiable and can reveal particularly sensitive details about individuals.	<ul style="list-style-type: none"> <li>• Addresses of incidents in 911 reports</li> <li>• Pickup and dropoff location of taxi trips</li> </ul>	Remove these fields or reduce the precision (i.e., generalize street address into zip code).
	Does the dataset contain geographic coordinates?	Although not human-interpretable, geographic coordinates can be easily mapped to a street address.	Geographic coordinates for the location of 311 requests	Does the dataset contain references to locations?



	ATTRIBUTE	RISK DESCRIPTION	EXAMPLE(S)	MITIGATING ACTION(S)
<b>Category 3: Sensitive fields and subsets</b>	Does the data contain any unstructured text fields?	Unstructured text fields are often used in unpredictable ways, meaning that their publication may expose unexpected sensitive information.	Permit applications that include the applicant's explanation of why the permit is required.	Remove the unstructured fields entirely or evaluate the entries to check for sensitive information.
	Does the data contain any types of records that are particularly sensitive?	Certain categories of records within a dataset may be systematically more sensitive than the rest.	Sexual assault incidents within a dataset of crime incident reports.	Treat these records with particular care, either by removing the entries entirely or removing/generalizing sensitive fields from these entries.
	Does the data contain information that also appears in other datasets?	Connecting information across multiple datasets may reveal sensitive information that is not contained within any individual dataset. This is known as the mosaic effect.	Demographic information (e.g., age, race, and gender) that appears in multiple datasets.	Remove or reduce the precision of any fields that are present in other public data.

One method that can be used to evaluate data for potential privacy concerns is to explicitly attempt to re-identify that data before it is shared. While it is perhaps counterintuitive to try to exploit vulnerabilities, it is better to identify risks internally rather than only after data is shared publicly. This approach is inspired by the idea of penetration testing, which was developed in the context of information security to test system robustness by explicitly trying to find and exploit vulnerabilities.<sup>68</sup> By performing penetration tests in-house, system managers in data security identify and remedy any issues before releasing software to the public.

A similar approach could be used to evaluate data for privacy risks before sharing it publicly. We call this “re-identification testing.” Open data leaders could work with internal data scientists or trusted community partners (such as members of the civic tech community) and have them try to re-identify individuals in certain datasets. While this process is more resource-intensive than following rules of thumb regarding privacy protection techniques, it is also far more robust at identifying privacy vulnerabilities in the data. By identifying the presence and types of privacy risks in a dataset, municipal data officers will be able to make more informed risk assessments and decisions regarding whether and how to release data.

One framework for re-identification testing is the “motivated intruder” model. This approach imagines the presence of an individual who is both technically competent and motivated to access personal information using the data. The Information Commissioner’s Office in the UK describes a motivated intruder as “a person who starts without any prior knowledge but who wishes to identify the individual from whose personal data the anonymised data has been derived,” and explains “This test is meant to assess whether the motivated intruder would be successful.”<sup>69</sup> The organization conducting a motivated intruder test assesses both the potential motivations of an intruder and the methods through which someone would go about achieving these aims. By focusing on an individual behind a re-identification attack rather than just the data in the abstract, the motivated intruder model provides a structured way to think through how a dataset might be used to attack individual privacy.

---

<sup>68</sup>Stephen Northcutt et al. “Penetration Testing: Assessing Your Overall Security Before Attackers Do.” (2006)

<sup>69</sup>Information Commissioner’s Office, “Anonymisation: managing data protection risk,” (2012).

## 2.4 DELETE: WHERE APPROPRIATE, RETIRE DATA STORED INTERNALLY, TURN OFF AUTOMATIC COLLECTION, AND REMOVE DATA SHARED ONLINE TO MITIGATE PRIVACY RISKS THAT RESULT FROM THE ACCUMULATION OF DATA.

As datasets grow and more datasets are published, it becomes easier to re-identify individuals and infer information about them. Deleting and unpublishing data are two strategies to limit the amount of information available and the concomitant risks to personal privacy.

The last stage of the data lifecycle is retiring data. This involves removing data from open data platforms as well as deleting internal records. At this stage, cities must balance the potential benefits of retaining data with the privacy risks that this data — and, in particular, the accumulation of many years of data — may trigger.

Although modern advances in data storage and data mining algorithms may suggest that more data is always better, Bruce Schneier asserts, “data is a toxic asset.”<sup>70</sup> When it comes to individual privacy, data often has negative retentive value: as datasets grow over time, they acquire more nuanced information about more people. The following table summarizes arguments for retaining and deleting data.

---

<sup>70</sup>Bruce Schneier, “Data is a toxic asset, so why not throw it out?,” CNN, March 1, 2016. <http://www.cnn.com/2016/03/01/opinions/data-is-a-toxic-asset-opinion-schneier/index.html>.

<sup>71</sup>de Montjoye et al., “Unique in the Crowd: The privacy bounds of human mobility.”

<sup>72</sup>de Montjoye et al., “Unique in the shopping mall: On the reidentifiability of credit card metadata.”

TOPIC	PRO: RETAIN DATA	CON: DELETE DATA
<b>Inference and analysis</b>	Historical data can become valuable when combined with other data or as algorithms develop new capabilities. Ten years of data might yield insights that one year of data would not. Historical data can also be valuable for cities trying to study trends or evaluate policies. Good data is critical for effective data-driven governance.	Larger quantities of data provide information about more people and yield more information about them. Studies have shown that the more data that exists the easier it is to identify individuals. <sup>71, 72</sup> Patterns of behavior that may not stand out in data about 1,000 taxi trips, for example, may be highly unique — and far more revealing — when 10,000 taxi trips are analyzed. This is especially concerning due to public records laws that limit a government's discretion over what data to release (see <a href="#">Section 3.4</a> ).
<b>Maintenance</b>	Increasing data storage capabilities make retaining and using data easy and inexpensive. It may take more work to evaluate and delete data than to simply leave it alone.	Collecting and maintaining more data than is needed requires the use of additional business and technical resources to manage inventories and workflows. Furthermore, the burden of responding to public disclosure requests increases with the volume of data maintained.

It is also important to recognize that open data, once released, is forever in the public domain. PCAST writes, “given the distributed and redundant nature of data storage, it is not even clear that data can be destroyed with any useful degree of assurance. Although research on data destruction is ongoing, it is a fundamental fact that at the moment that data are displayed (in ‘analog’) to a user’s eyeballs or ears, they can also be copied (‘re-digitized’) without any technical protections.”<sup>73</sup> Thus, even if a city removes a sensitive dataset from its portal, that data is likely to remain available online through channels such as archives and peer-to-peer networks. For example, in 2006 AOL released a dataset of 20 million web searches by 650,000 Americans. Even though this data was immediately exposed as highly sensitive and AOL quickly removed the data from its own site, to this day the data remains accessible online through a quick web search.<sup>74, 75</sup>

<sup>73</sup>President’s Council of Advisors on Science and Technology, “Big Data and Privacy: A Technological Perspective.”

<sup>74</sup>Michael Barbaro and Tom Zeller Jr., “A Face Is Exposed for AOL Searcher No. 4417749,” The New York Times, August 9, 2006. [www.nytimes.com/2006/08/09/technology/09aol.html](http://www.nytimes.com/2006/08/09/technology/09aol.html)

<sup>75</sup>Greg Sadetsky, “AOL search data mirrors,” [http://www.gregsadetsky.com/\\_aol-data/](http://www.gregsadetsky.com/_aol-data/)

# TAKE ACTION

Determining whether to retain data or turn off automatic collection processes requires proactive planning and active balancing of potential benefits and risks. For data that should not be retained or published beyond a predetermined period, cities should build automatic deletion into the technology architecture. Cities should also routinely review their automatic collection processes. When deciding to retain data (beyond required retention periods), cities should be confident that the data will produce enough value to justify any privacy risks. The worst-case scenario is that retained data will produce no positive value but reveal sensitive information about citizens. If data is particularly sensitive, there must be valuable uses planned that justify the retention risks.

Cities should follow the following three-step process to determine the value of retaining data:

## 1. Determine benefits

Past value

Future value		<b>LOW:</b> The data has not been used	<b>MEDIUM:</b> The data is used to create mild impact	<b>HIGH:</b> The data has been used to create high impact
	<b>LOW:</b> There are no plans to use the data	Low benefit		
	<b>MEDIUM:</b> There are tentative plans to use the data to create mild impact		Medium benefit	
	<b>HIGH:</b> There are clear plans to use the data to create high impact			High benefit

## 2. Determine risks

Impact

Scale		<b>LOW:</b> Minimally sensitive information (or mildly sensitive information that is not public record)	<b>MEDIUM:</b> Mildly sensitive information that is public record	<b>HIGH:</b> Very sensitive information
	<b>LOW:</b> The data does not become more sensitive as it grows	Low risk		
	<b>MEDIUM:</b> The data may become more sensitive as it grows		Medium risk	
	<b>HIGH:</b> The data becomes more sensitive as it grows			High risk

### 3. Balance benefits and risks

Benefit

Risk		LOW	MEDIUM	HIGH
	LOW			Retain data or continue collection
	MEDIUM		Make determination based on the particular risks and potential for effective mitigation	
	HIGH	Delete data or stop collection		

If the decision is made to remove existing open data from an online portal, it is important to:

- Determine who uses the data, and communicate with any businesses, organizations, or individuals that rely on the data.
- Explain to the public why data is being removed, making it clear that data is being removed to protect privacy rather than to subvert transparency, and clarify the timeline for this process.



# IN PRACTICE

In order to improve its ability to recover stolen cars, the Seattle Police Department (SPD) implemented Automatic License Plate Recognition (ALPR) technology on several of its patrol cars. When these patrol cars were in service, the ALPRs read license plates that came in view and associated it with GPS location data. Images of these plates were then translated into text data, which was stored on an internal server along with the location at which that license plate had appeared, to be cross-referenced whenever a vehicle was reported as stolen.

When the ACLU submitted a public records request in 2011 for data from the SPD's license plate reader program, they found that the City possessed multiple years worth of license plate scans: their database contained 7.3 million records of license plates and locations from a three-year period.<sup>76</sup> This was especially concerning given that only 7,244 (0.1%) of the records contained hits on stolen vehicles. The ACLU mined this database to learn behavioral patterns and location history of officers (only to demonstrate the sensitivity of this information), and presumed that it could do much of the same for the people whose cars had been scanned. Given that most people are identifiable from just a few spatiotemporal (location and time) data points of behavior,<sup>77, 78</sup> the data collected by the SPD — and available to anyone who submitted a public records request — could reveal sensitive behavioral patterns about many individuals.

Because the SPD had not developed a retention policy at the point of the ACLU's public records request, it had simply stored all of the records even beyond the time that the data was being used. The vast quantity of data meant that it was possible to learn about more people and with a greater level of detail than if only the most recent records had been available. After this incident, retention policies for license plate data were changed to require that records be kept for no more than 90 days.<sup>79</sup>

---

<sup>76</sup>Jamela Debelak, "ALPR: The Surveillance Tool You've Probably Never Heard Of," May 20, 2013. <https://aclu-wa.org/blog/alpr-surveillance-tool-you-ve-probably-never-heard>

<sup>77</sup>de Montjoye et al., "Unique in the shopping mall: On the reidentifiability of credit card metadata."

<sup>78</sup>de Montjoye et al., "Unique in the Crowd: The privacy bounds of human mobility."

<sup>79</sup>Brian M. Rosenthal, "Police cameras busy snapping license plates," The Seattle Times, August 3, 2013. [www.seattletimes.com/seattle-news/police-cameras-busy-snapping-license-plates/](http://www.seattletimes.com/seattle-news/police-cameras-busy-snapping-license-plates/)

### 3. DEVELOP OPERATIONAL STRUCTURES AND PROCESSES THAT CODIFY PRIVACY MANAGEMENT WIDELY THROUGHOUT THE CITY.

Protecting privacy is the responsibility of everyone in a city who collects, manages, or uses data. Thorough organizational structures and processes are robust ways to ensure that privacy is managed responsibly.

Cities must develop rigorous and robust approaches to managing privacy in open data. Although privacy management is often seen as restrictive, it is a fundamental component of innovative and data-driven policy. Because many of the most ambitious efforts in government involve the use of data, they will not be possible without effective privacy management. Marc Groman, Senior Adviser for Privacy at the Office of Management and Budget, describes privacy's role as follows:

"if you have a well-resourced, well-functioning privacy program, that program will promote innovation, will actually foster and enable more information sharing and allow agencies to expand to new technologies such as the cloud or mobile [...] if you have the right talent in a privacy program, then all initiatives and the agency's mission ultimately will be improved and will be able to accomplish more things and better things for the American people."<sup>80</sup>

In other words, effective privacy management is essential to maximizing the impact of open data. Without the ability to manage and mitigate privacy risks, open data programs and the city as a whole will not be able to realize the full potential of collecting, using, and sharing its data.

Building effective privacy management requires an internal appreciation for the value and role of privacy. While all cities are familiar with the related area of data security, privacy is a relatively new area of focus. Although privacy and security are related, Derek Bambauer explains that they "are distinct concerns. Privacy establishes a normative framework for deciding who should legitimately have the capability to access and alter information. Security implements those choices."<sup>81</sup> The field of information security has developed process-based standards that emphasize best practices to minimize the risks of breaches, and evaluates efforts based on their adherence to this process rather than on particular outcomes. Drawing on this tradition in security, and given that it is impossible to guarantee perfect data anonymity, Ira Rubinstein and Woodrow Hartzog write, "the best way to move data release policy past the alleged failures of anonymization is to focus on the process of

---

<sup>80</sup>Aitoro, "Defining privacy protection by acknowledging what it's not."

<sup>81</sup>Derek E Bambauer, "Privacy Versus Security," *Journal of Criminal Law and Criminology* 103, no. 3 (2013).

minimizing risk of reidentification and sensitive attribute disclosure, not preventing harm. Process-based data release policy, which resembles the law of data security, will help us move past the limitations of focusing on whether data sets have been “anonymized.”<sup>82</sup>

Effective open data management must therefore be operationalized at the process level in addition to the data level. This Chapter focuses on strategies and processes to institutionalize the privacy management practices described in [Chapter 1](#) and [Chapter 2](#).

---

<sup>82</sup>Rubinstein and Hartzog, “Anonymization and Risk.”

## 3.1 INCREASE INTERNAL AWARENESS OF AND ATTENTION TO PRIVACY RISKS.

The responsibility for managing privacy in open data spans every office in city hall. Assembling privacy-conscious teams and processes is critical for managing diverse privacy risks.

As described in [Chapter 1](#), the privacy risks of open data are diverse. The breakdown of the PII framework means that there are no simple, formulaic answers for when data is sensitive nor easy answers that will ensure data can be safely released. Data can be sensitive in ways that are unexpected to those who are unfamiliar with re-identification. On the other hand, that some risks to privacy will inevitably exist does not mean that cities should simply stop releasing data.

Furthermore, [Chapter 2](#) explains how privacy is a critical consideration at all stages of the data lifecycle: when data is collected, as data is managed, when datasets are released, and as data is deleted. Every group that collects, uses, or shares data makes decisions that implicate privacy; many such decisions are made throughout city hall every day.

The scale and complexity of managing privacy makes it infeasible for cities to ignore privacy considerations or assign all responsibility to a single individual. Instead, effectively managing privacy requires a diverse team that collectively spans city hall and is capable of considering privacy in connection with every decision regarding data. Furthermore, due to the diversity of privacy risks, groups responsible for privacy must work together to avoid sensitive information falling through the cracks.

# TAKE ACTION

Cities should develop comprehensive systems and processes, along with tailored training for each role, to manage privacy. The following table describes the key positions that every city should have, and their primary roles and responsibilities. Along with others such as community members, these people should also form a citywide Privacy Review Board (inspired by Institutional Review Boards, or IRBs, that oversee studies on humans to ensure they are conducted legally and ethically).

TITLE	ROLE <i>What is their general mission?</i>	TRAINING <i>What do they need to know?</i>	DATA RELEASE PROCESS RESPONSIBILITIES <i>What is their role in releasing open data?</i>
<b>Chief Privacy Officer</b>	<ul style="list-style-type: none"> <li>• Lead citywide efforts to manage and protect privacy</li> <li>• Emphasize that privacy is an integral part of open data, not just a hurdle.</li> </ul>	<ul style="list-style-type: none"> <li>• How to manage a privacy-conscious organization.</li> <li>• How to engage the public about data and privacy.</li> </ul>	<ul style="list-style-type: none"> <li>• Manage data preparation for particularly challenging datasets.</li> <li>• Lead regular risk assessments of open data datasets and processes.</li> </ul>
<b>Open Data Team</b>	<ul style="list-style-type: none"> <li>• Determine how data is released.</li> <li>• Evaluate the ecosystem of data available on the portal.</li> <li>• Advise departments how to prepare sensitive data for release.</li> </ul>	<ul style="list-style-type: none"> <li>• The diversity of re-identification risks and re-identification techniques.</li> <li>• How to mitigate re-identification risks to make datasets appropriate for public consumption.</li> <li>• Best practices for managing public relations regarding data privacy.</li> </ul>	<ul style="list-style-type: none"> <li>• Work with Department Privacy Officers to perform privacy risk assessments on data being considered for release.</li> <li>• Ensure that datasets being released have thorough metadata that describe any protections taken to protect privacy in the data.</li> </ul>
<b>Department Data or Privacy Officer</b>	<ul style="list-style-type: none"> <li>• Manage data and privacy within a particular department or small number of departments.</li> </ul>	<ul style="list-style-type: none"> <li>• Fluency with the types of data used by that officer's department(s).</li> <li>• Best practices for managing privacy throughout the data lifecycle.</li> <li>• Overview of re-identification risks and mitigations.</li> </ul>	<ul style="list-style-type: none"> <li>• Curate datasets within departments.</li> <li>• Provide comprehensive metadata on datasets being considered for release.</li> <li>• Perform risk assessment on departmental data to determine how it might reveal insights about individuals.</li> </ul>
<b>City Attorney</b>	<ul style="list-style-type: none"> <li>• Advise individuals and groups throughout the City regarding the laws behind data management and privacy.</li> </ul>	<ul style="list-style-type: none"> <li>• Pertinent privacy and open records laws.</li> <li>• Potential liabilities for releasing data.</li> <li>• Best practices in privacy management.</li> </ul>	<ul style="list-style-type: none"> <li>• Identify if the data (in whole or in part) is public record.</li> <li>• Determine if any of the dataset is affected by privacy laws.</li> </ul>

# IN PRACTICE

As part of its Privacy Program, the City of Seattle released in February 2016 an Open Data Policy that defined a variety of roles involved in curating, releasing, and maintaining open data.<sup>83</sup> These positions are defined in the Policy as follows:

TITLE	DESCRIPTION
<b>Open Data Manager</b>	A City employee who is responsible for the City of Seattle's Open Data Program, stewards the data made available on the Open Data Portal, and manages the Open Data Team.
<b>Open Data Team</b>	City employees who administer the Open Data Portal and provide planning, review, coordination, and technical support to City departments and offices publishing open data.
<b>Data Owner</b>	Any City employee who is responsible for one or more departmental datasets.
<b>Open Data Champion</b>	Designated by each department, this person serves as the point of contact and coordinator for that department's publishing of Open Data.
<b>Chief Privacy Officer</b>	A City employee who provides overall leadership and direction to the City of Seattle Privacy Program and is responsible for resolving questions and issues concerning privacy risk and Open Data.

Seattle's policy also dictates the responsibilities of each position. For example, the Chief Privacy Officers' responsibilities include:<sup>84</sup>

- "Serve as the arbiter of any questions or issues concerning Open Data privacy risk and solutions for minimizing the risk of privacy harm."
- "Resolve questions or issues regarding open data privacy risk escalated by the Open Data Manager."
- "Participate in the annual risk assessment for the Open Data Program" (see [Section 3.2](#)).

The City of Boston maintains similar roles for its open data program. In order to help its open data program scale, in July 2016 the City of Boston reformed its process for how data is reviewed before being published. While Boston's open data portal had been operational since 2012, the City for many years had no formal process dictating how to prepare data for publication. In fact, the open data program manager often remarked that he did not know who had collected and uploaded a particular dataset, or why. This meant that there was no comprehensive inventory of the City's open data nor

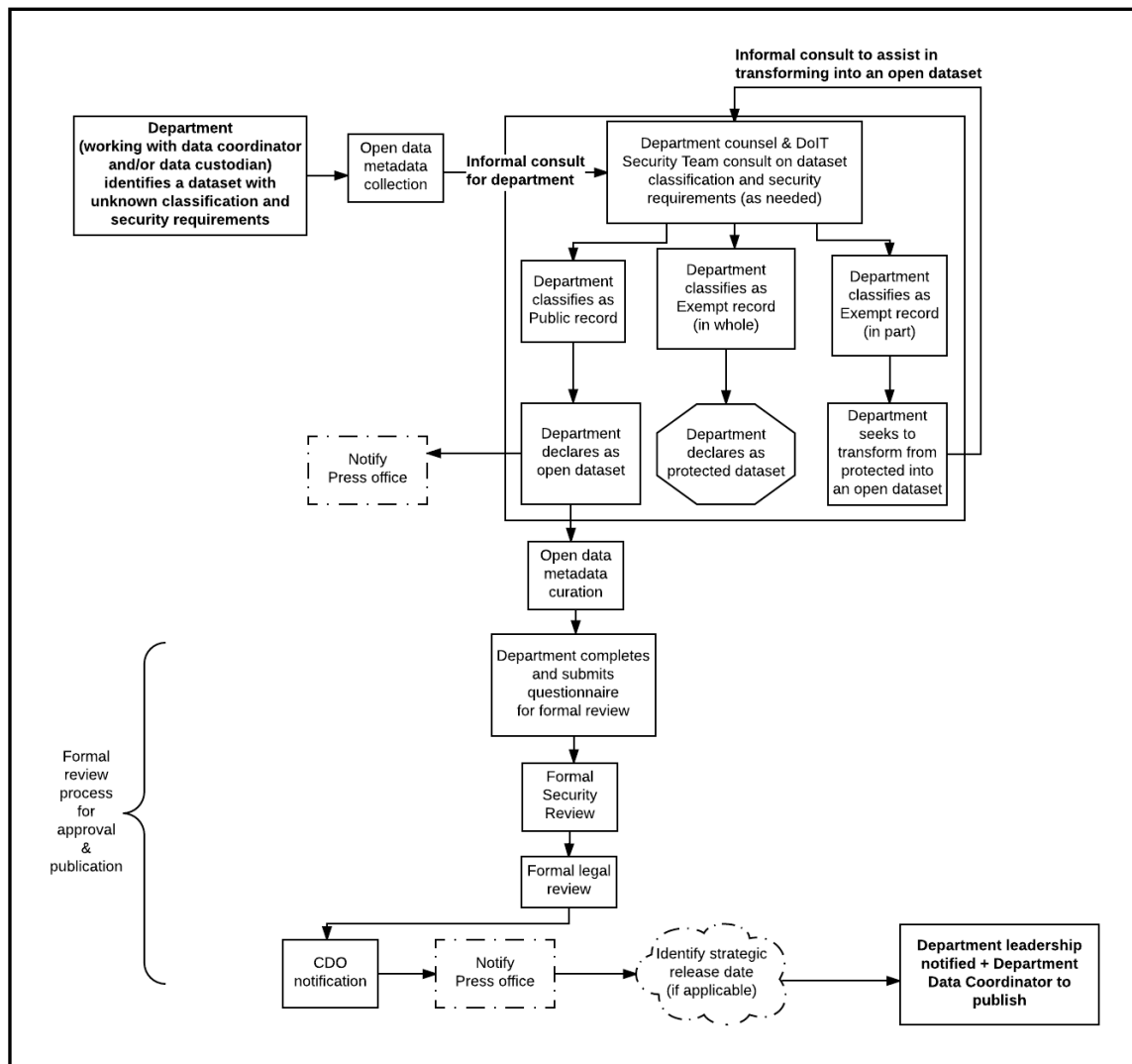
---

<sup>83</sup>City of Seattle. "Open Data Policy." (2016) <http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPolicyV1.pdf>.

<sup>84</sup>Ibid.

any ability to understand previous open data decisions. Given the many different groups involved in the publishing process, it was critical to develop procedures that clarified everyone's roles. The final review process (Figure 6) provides clear guidance as to how various City stakeholders should collaborate to ensure that data is properly vetted and prepared before publishing.

Figure 6. City of Boston dataset approval and publication process



## 3.2 PERIODICALLY AUDIT DATA AND PROCESSES TO ENSURE PRIVACY STANDARDS CONTINUE TO BE UPHELD.

As more data is released and the privacy landscape evolves, previously acceptable data practices may no longer be appropriate. Regularly evaluating approaches to data privacy will help characterize and mitigate any privacy risks that have emerged.

The rapidly shifting landscape of Big Data and privacy means that it is impossible to future-proof datasets and data release processes. Constant changes in technology, policy, and public perceptions regularly challenge previously held assumptions about what data is safe to release. Such changes can come from many directions, such as:

- More powerful re-identification techniques that emerge over time increase what can be inferred from a dataset, potentially rendering earlier mitigation approaches untenable.
- New datasets (on the open data portal and elsewhere) increase the potential for re-identification through the mosaic effect (i.e., linking information across datasets), and increase the potential negative impact of such re-identifications.
- Personnel turnover and changes in internal structures may diminish the institutional capacity and knowledge needed to effectively manage data privacy.
- Developments in the national and local discourse about privacy — perhaps resulting from privacy scandals related to the government or a private company — can alter the public's faith in government as a data steward and willingness to accept privacy risk as a cost of open data.



# TAKE ACTION

Cities should undergo regular audits (e.g., annually) to ensure that their data and processes achieve desired levels of privacy protection, and should document the results of these audits. This requires evaluating the program's past performance in protecting privacy and any broad shifts that have occurred in the data and privacy landscape. The answers to these questions will help cities evaluate their open data program's performance and identify necessary changes.

	<b>PAST PERFORMANCE</b> <i>How well has the city managed open data and privacy in the past?</i>	<b>LANDSCAPE SHIFTS</b> <i>What recent developments require updated processes?</i>
<b>Data and algorithms</b>	<ul style="list-style-type: none"> <li>• What is the current risk assessment of every data-set available?</li> <li>• How do today's risk assessments compare with those from when the data was released?</li> <li>• What re-identifications have occurred based on the available data?</li> </ul>	<ul style="list-style-type: none"> <li>• What new datasets have been made available since the last audit?</li> <li>• What are the new forms of re-identification attacks that have emerged since the last audit?</li> <li>• What inferences are possible by combining all the data on the portal?</li> <li>• What new types of data have been made public beyond our open data portal (or are collected by private actors) that could be linked with the data we have made available?</li> <li>• How does the data we have released compare with those available from other cities?</li> </ul>
<b>Process</b>	<ul style="list-style-type: none"> <li>• How effective is our process for releasing data?</li> <li>• Is it timely?</li> <li>• Is every task covered by someone with the appropriate expertise?</li> <li>• Has sensitive information been inadvertently released? If so, how did it happen?</li> </ul>	<ul style="list-style-type: none"> <li>• How has the organizational structure of City Hall changed?</li> <li>• What shifts in personnel have occurred?</li> <li>• What gaps in expertise and process have been emerged?</li> </ul>
<b>Public</b>	<ul style="list-style-type: none"> <li>• How effective has our public engagement been? (see <a href="#">Chapter 4</a>)</li> <li>• Have there been any incidents of negative feedback from the public regarding privacy?</li> </ul>	<ul style="list-style-type: none"> <li>• How have public perceptions of privacy shifted since our last audit?</li> <li>• What issues of privacy have dominated the news?</li> <li>• Have there been any privacy scandals (related to our city, neighboring jurisdictions, or state or federal governments) that could influence public trust related to data?</li> <li>• What best practices have emerged in other cities?</li> </ul>

# IN PRACTICE

In 2016, Seattle committed to performing an annual risk assessment of its open data ecosystem. The City's Open Data Policy states that it will "Perform an annual risk assessment of both the Open Data Program and the content available on the Open Data Portal. The outcome of this review shall be shared with the Open Data Champions, who will help implement risk mitigation strategies."<sup>85</sup>

Seattle's initial risk assessment is being conducted in partnership with the Future of Privacy Forum (FPF). FPF will evaluate Seattle's existing open data inventory, and will also spend time in City Hall to gauge the efficacy of Seattle's approaches and processes for managing data privacy. An excerpt from FPF's Statement of Work is shown below.

## FPF'S STATEMENT OF WORK

FPF will produce a public-facing report on the City of Seattle's Open Data program. The risk assessment methodology and report will be based on review of a subset of high-risk agencies or data sets, as well as a random sample of additional agencies or data sets. FPF will evaluate the process in place to determine the risk of re-identification in case of release of individual data sets or multiple data sets. Within this scope of review, FPF will

- Assess the municipal open data landscape and identify inherent risks.
- Assess the City of Seattle as a model municipality, its organizational structure and data handling practices related to Open Data, and identify potential risks.
- Identify data privacy, security, and handling risks as they apply to an open data program.
- Develop a template/framework that facilitates the scoring of risk in the areas identified.
- Apply the framework through an assessment of the City's current Open Data Program, any mitigating activities, and measure residual risk.
- Propose recommendations to mitigate risk to acceptable levels.
- Issue a report that will be shared publicly with the Seattle community.
- Share the risk management methodology, tools, and other collateral publicly to benefit other municipalities.
- Promote the report and project through submissions for IAPP conferences and publications.

Where possible, the risk assessment methodology and related tools and controls will be based on an open framework such as the National Institute of Standards and Technology (NIST) Special Publication 800 series (e.g. 800-30, 800-37, 800-53).

---

<sup>85</sup>Ibid.

<sup>86</sup>Future of Privacy Forum, "Future of Privacy Forum," <https://fpf.org>

### 3.3 ACCOUNT FOR THE UNIQUE RISKS AND OPPORTUNITIES PRESENTED BY PUBLIC RECORDS LAWS.

Public records laws compel governments to release data requested by citizens, meaning that even when a city might not have proactively published data, information may still be made public through a different channel. Cities should therefore limit the data they collect and store, while also tapping into public records requests and coordinating public records and open data processes so as to better understand and serve public interest in data.

Public records laws refer to statutes stipulating that certain government data, when it has been requested by a member of the public, must be released. The most well-known of these statutes is the Freedom of Information Act, or FOIA, that dictates the release of information from the Federal Government unless a particular exception bars its release; every state also has a public records law. These laws are designed to promote transparency in government by guaranteeing public access to government information. Yet while the release of public records is an effective tool to enhance government transparency, such releases involve making data public and therefore also have important implications with regard to privacy. Coordination is essential.

Although public records laws do not dictate open data policies, there are strong connections between the two given that both involve sharing government data with the public. Whereas open data is proactive, meaning that cities choose what data to release, public records are inherently reactive: cities must share whatever information is requested (barring exemptions, see below). This means that whereas cities have discretion over how to publicly share data on open data portals, public records laws often force cities' hands into releasing information that might implicate individual privacy. In other words, while all open data is public record, not all public records should be open data.

Public records laws typically contain a number of exemptions setting forth categories within which government data is not required to be released even when requested. For example, the Massachusetts public records law has over 20 such exemptions, covering reasons such as statutory restrictions, law enforcement investigations, and privacy.<sup>87</sup> Yet an official guide to the law (and applicable case law) emphasizes that these exemptions should be "narrowly construed" in order to serve overall transparency goals, meaning that exemptions should limit the release of public records only in specific situations.<sup>88</sup> The privacy exemption itself is quite specific, covering

only “personnel and medical files or information; also any other materials or data relating to a specifically named individual, the disclosure of which may constitute an unwarranted invasion of personal privacy.”<sup>87</sup> The strictness of the privacy exemption means that only “intimate details of a highly personal nature” (such as substance abuse, government assistance, and family reputation) data could be withheld from public disclosure due to privacy considerations.<sup>90</sup> Much of the information that creates risks in open data would not be exempt. Thus, while public records laws contain exemptions, these are not reliable protectors of individual privacy.

---

<sup>87</sup>Massachusetts Public Records Definition. Mass. Gen. Laws ch. 4, § 7(26).

<sup>88</sup>William Francis Galvin. “A Guide to the Massachusetts Public Records Law.” (2013) <http://www.sec.state.ma.us/pre/prepdf/guide.pdf>.

<sup>89</sup>Mass. Gen. Laws ch. 4, § 7(26).

<sup>90</sup>Attorney General v. Assistant Commissioner of the Real Property Department of Boston, 380 623 (1980).

# TAKE ACTION

Cities should always be mindful of the implications public records laws may have for them. Below are several actions that cities can follow to take advantage of public records laws while mitigating the risks they pose.

ACTION	DESCRIPTION
<b>Limit collection of sensitive data that would be public record.</b>	Data that is public record can be released into the public domain whether or not it is proactively released as open data: the public can compel the government to release this information by filing a public records request. Before collecting new data, therefore, it is critical to determine what information would be a public record. If this data is sensitive, it is often best not to collect it in the first place as it otherwise could be released whenever requested (see <a href="#">Section 2.1</a> for guidelines on limiting data collection).
<b>Delete sensitive information that is public record.</b>	When maintaining data, audit existing inventories to determine what sensitive information is public record. Given the potential for public disclosure of sensitive data at any time, delete data that is no longer generating enough value to merit the privacy risk (assuming that it has been stored for the mandated retention period; see <a href="#">Section 2.4</a> for guidelines on deleting data).
<b>Release frequently requested information as public records.</b>	When deciding what datasets are good candidates to become open data, consider those that are often subject to public records requests. Such datasets are of clear public interest. Furthermore, releasing these datasets will reduce the burden of fulfilling public records requests: a recent study of FOIL requests to New York State's Department of Environmental Conservation found that the Department "could reduce FOIL requests by 50% by publishing frequently FOILed data." <sup>91</sup>

---

<sup>91</sup>Reinvent Albany. "Listening to FOIL: Using FOIL Logs to Guide the Publication of Open Data." (2014)

One example of sensitive data being released through public records requests comes from the case study presented in [Section 1.3](#). To recap: in response to a Freedom of Information Law request, New York City released data detailing every taxi ride recorded in registered NYC taxis in 2013.<sup>92</sup> The data contained information about pickup time and location, dropoff time and locations, and the taxi cab and driver involved in each trip. As described in [Section 1.3](#), one data scientist used this information to re-identify the drivers involved in every trip.

Another data scientist, Anthony Tockar, took the taxi data and analyzed the patterns of trips. To show how the trips could reveal sensitive information beyond the identities of taxi drivers, he focused on studying the behavioral trends of visitors to a strip club in New York City.<sup>93</sup> Tockar mapped out the dropoff location of every taxi trip that began outside a particular strip club in Hell's Kitchen between midnight and 6am. By looking for locations at which many taxi trips ended, Tockar was able to identify where frequent visitors lived. Due to the precision of the geographic coordinates provided, Tockar "was able to pinpoint certain individuals with high probability."<sup>94</sup> Combining this data with auxiliary information available online enabled easy re-identification. Tockar writes,

Examining one of the clusters in the map above revealed that only one of the 5 likely drop-off addresses was inhabited; a search for that address revealed its resident's name. In addition, by examining other drop-offs at this address, I found that this gentleman also frequented such establishments as "Rick's Cabaret" and "Flashdancers". Using websites like Spokeo and Facebook, I was also able to find out his property value, ethnicity, relationship status, court records and even a profile picture!"<sup>95</sup>

That such information was made available through public records requests indicates that non-exempt government information can contain sensitive information about individuals. The privacy exemptions in public records laws are defined following the traditional PII framework, and are therefore not designed to protect against re-identification related to the quasi identifiers in data such as taxi trips. It is therefore critical to be diligent when releasing open data about privacy above and beyond the particular privacy exemptions within public records laws. Open data officials need to work closely with individuals responsible for responding to public records requests, and cities should require these efforts be coordinated to the maximum extent possible.

---

<sup>92</sup>Whong, "FOILing NYC's Taxi Trip Data".

<sup>93</sup>Tockar, "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset".

<sup>94</sup>Ibid.

<sup>95</sup>Ibid.



## 3.4 DEVELOP A PORTFOLIO OF APPROACHES FOR RELEASING AND SHARING DATA.

Publicly releasing full datasets is not always the best strategy for capturing benefits while mitigating risks. Developing and using a wide range of data-sharing strategies will help ensure that data is shared responsibly and effectively.

Most discussions of open data emphasize the full release of raw, granular data. Yet although there are many benefits to sharing government data, the privacy risks involved mean that it does not always make sense to make a raw dataset publicly available to anyone and everyone.

Nonetheless, datasets that are too sensitive to fully open up are often those with the most potential value for novel analyses and impactful applications. These datasets must move beyond the walls of city halls without becoming truly open to the entire public. This means that cities must deploy tiered-access structures that allow for carefully tailored levels of data openness.

The Open Data institute<sup>96</sup> has developed a Data Spectrum that distinguishes openness across the five following categories:<sup>97,98</sup>

- Closed data: “Data that can only be accessed by its subject, owner or holder.”
- Shared data
  - Named access: “data that is shared only with named people or organisations”
  - Attribute-based access: “data that available to specific groups who meet certain criteria”
  - Public access: “data that is available to anyone under terms and conditions that are not ‘open’”
- Open data: “Data that anyone can access, use, and share”.

The Data Spectrum framework highlights that there are many options available for cities to share data that go beyond the binary notion of releasing or withholding information.

---

<sup>96</sup>Open Data Institute, “Open Data Institute,” <http://theodi.org>.

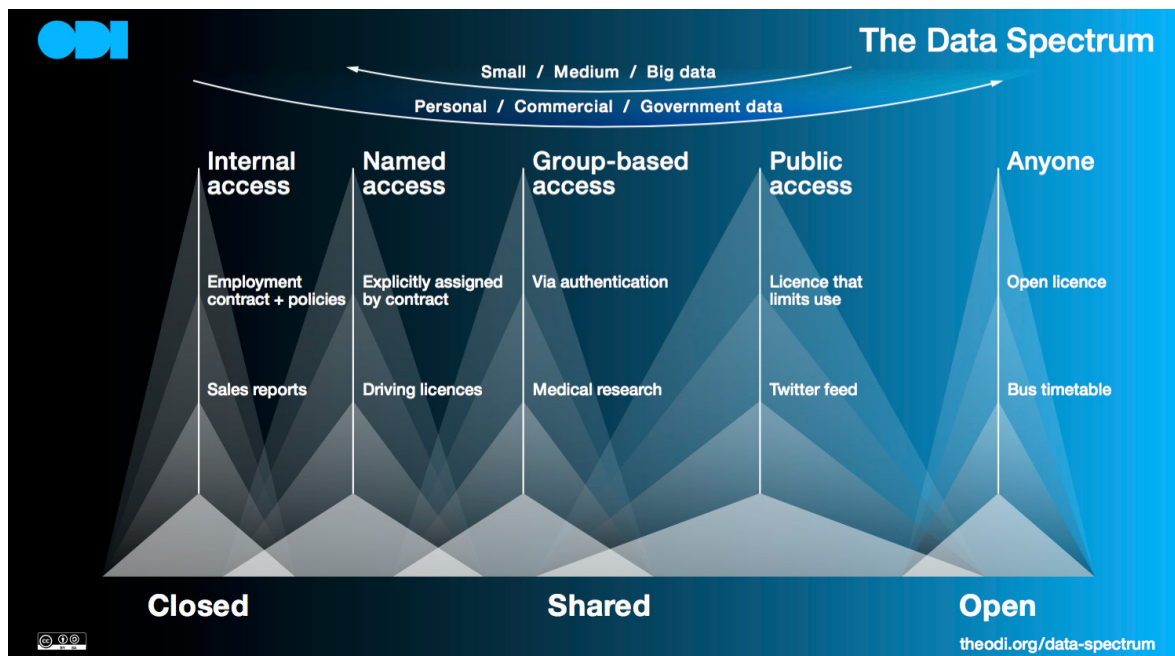
<sup>97</sup>“The Data Spectrum,” <http://theodi.org/data-spectrum>.

<sup>98</sup>Ellen Broad, “Closed, shared, open data: what’s in a name?,” <https://theodi.org/blog/closed-shared-open-data-whats-in-a-name>.



By enabling cities to share data with researchers, government agencies, and other trusted organizations, the five tiers of shared data listed by the Open Data Institute are critical for maximizing the possibilities of municipal data. Since trusted groups are among the most likely to use open data, sharing with only them can allow cities to capture most of the benefits that fully open data would realize without exposing people's privacy to the same risks.

Figure 8. Open data spectrum



From The Data Spectrum, <http://theodi.org/data-spectrum>.

<sup>99</sup>"BOS:311," <https://311.boston.gov>

<sup>100</sup>City of Boston, "311, Service Requests," <https://data.cityofboston.gov/City-Services/311-Service-Requests/awu8-dc52>.

<sup>101</sup>O'Brien, Gordon, and Baldwin, "Caring about the community, counteracting disorder: 311 reports of public issues as expressions of territoriality,"; Daniel Tumminelli O'Brien, Robert J Sampson, and Christopher Winship, "Econometrics in the Age of Big Data: Measuring and Assessing "Broken Windows" Using Large-scale Administrative Records," *Sociological Methodology* 45 (2015).

# TAKE ACTION

Cities should consider the results of the risk assessments from [Chapter 1](#) when determining how to share data as well as whether to share it. Data that a) has large potential benefits if shared, but b) involve particularly sensitive information, are good candidates for the intermediate levels of the Open Data Spectrum described above. For these datasets, contractual arrangements allow cities to leverage partners such as academic researchers and other government agencies while minimizing the privacy risks to the individuals whose data is being shared.

To improve the process of sharing data, open data programs should:

- Develop model Data Use Agreements and Memorandums of Understanding that can be tailored for many datasets and partners. These agreements can ensure proper use of sensitive data by explicitly limiting the uses of the data (such as using the data to re-identify individuals), identifying the party responsible for maintaining confidentiality, establishing policies regarding reuse of the data, and providing a structure for penalties if data is used inappropriately.
- Require that researchers requesting data beyond what is appropriate for open data provide clear justifications for the data they desire. Even though the data will be provided with restrictions on sharing and uses, it is still best practice to avoid unnecessarily sharing sensitive information.
- Consider the social status of the external partner to ensure public support for sharing the data. The public may be more comfortable with data being shared with another governmental agency to benefit the community than with a private company to help them profit.
- Document the steps taken to enable and ensure tiered access.

---

<sup>99</sup><https://311.boston.gov>

Since 2009, The City of Boston has maintained a smartphone application to support constituent requests for service through its 311 system. When someone submits a 311 request through the app, their request and location are recorded, along with (if submitted) the user's name, user ID, and email. Boston makes some information from these constituent requests available on its open data portal. Due to privacy concerns, the open dataset does not include personal information about the individuals who submitted requests. This is sensible, given that such data could be used to identify individuals and learn about their lives.

In addition to sharing a subset of data through its open data portal, Boston also shares the raw dataset with researchers through data sharing agreements, which provide full access to the data but prohibit sharing it with anyone not a party to the agreement. Such partnerships have led to numerous analyses and research papers, with insights related to the motivations and behaviors of those who report issues. The lessons learned from this research have fed back into the development of future versions of the 311 system, thereby making the data sharing agreements an integral component of improving the City's ability to deliver services.

---

<sup>100</sup>City of Boston, "311, Service Requests," <https://data.cityofboston.gov/City-Services/311-Service-Requests/awu8-dc52>.

<sup>101</sup>O'Brien, Gordon, and Baldwin, "Caring about the community, counteracting disorder: 311 reports of public issues as expressions of territoriality."; Daniel Tumminelli O'Brien, Robert J Sampson, and Christopher Winship, "Econometrics in the Age of Big Data: Measuring and Assessing "Broken Windows" Using Large-scale Administrative Records," *Sociological Methodology* 45 (2015).

## 4. EMPHASIZE PUBLIC ENGAGEMENT AND PUBLIC PRIORITIES AS ESSENTIAL ASPECTS OF DATA MANAGEMENT PROGRAMS.

Public support is a prerequisite for successful open data programs. Engaging the public in the development of policies and practices will build vital support and drive open data forward.

Open data is about more than just releasing datasets: it is, as described in an Executive Order signed by Barack Obama on his first day of office, essential “to ensure the public trust and establish a system of transparency, public participation, and collaboration.”<sup>102</sup> While open data initiatives often focus on the amount of data made available, Obama’s message is that open data is about more than just the data. Recognizing these broader goals allows us to situate open data in its proper context: open data is a means toward transparency and accountability, not the end.

This means that open data initiatives should evaluate their efforts based on the extent to which their process of releasing data achieves the underlying goals of transparency and accountability. Just like any government program, open data initiatives rely on a series of decisions balancing complex and often competing factors. These include what data to release, how to release it, and how to make the data valuable. Instead of viewing these challenges as roadblocks hindering their progress and requiring opaque decisions, leaders should recognize that open data is designed to promote transparency and embrace these challenges as opportunities for collaborative decision-making with constituents. What better place to pioneer open government processes than a program already designed precisely to promote open government?

As the rest of this chapter will show, effective public engagement should be an essential component of any open data program. This is especially true when it comes to privacy: since it is generally the public’s own privacy at stake, it is essential that citizens have a voice as to how data is released. This sense of agency is likely to generate trust in the cities ultimately making decisions. By following the principles described in this chapter, open data programs will promote open government far more than they would by merely releasing more datasets.

---

<sup>102</sup>Barack Obama, “Transparency and Open Government,” *The White House* 2009. [https://www.whitehouse.gov/the\\_press\\_office/TransparencyandOpenGovernment](https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment).

## 4.1 GARNER SUPPORT FOR OPEN DATA BY SHARING THE BENEFITS AND SUCCESSFUL USES OF OPEN DATA.

The public will not accept the privacy risks of open data unless they also recognize the value of sharing that data. Building and celebrating success encourages more citizens to support open data.

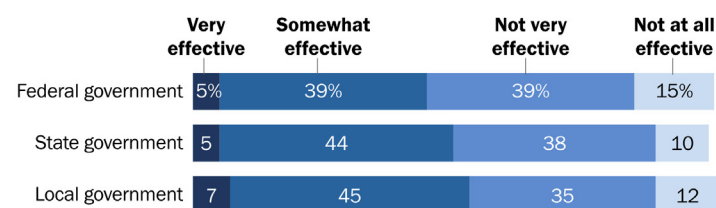
A recent Pew Research Center study found that cities have a long way to go to convince the public that open data releases are effective.<sup>103</sup> Only 7% of Americans feel that local governments are “very effective” in sharing data they collect with the public. This number is even lower when it comes to public perception of federal government data releases. Citizens also broadly doubt that these data disclosures have made government more effective.

A more nuanced issue rests in the commonly held belief that government open data disclosures currently do little to improve government performance. The same Pew study referenced above found that only 19% of Americans could think of an example where local government had provided useful information about data it collects, and just 9% of Americans think that “the data government shares with the public helps a lot with the private sector’s creation of new products and services.”<sup>104</sup> Ultimately, the public is split on whether open data “improves the quality of government services.” This perception need not continue to typify engagement with open data, however, as 66% of Americans are optimistic that open data may improve government accountability.<sup>105</sup>

Figure 9.

### Few Think Government at Any Level Shares Its Data Very Effectively

% of adults who judge the effectiveness of government data sharing to be ...



Online survey of 3,212 adults in Pew Research Center's American Trends Panel, Nov. 17-Dec. 15, 2014.

PEW RESEARCH CENTER

From Pew Research Center. “Americans’ Views on Open Government Data.” (2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

<sup>103</sup>Pew Research Center. “Americans’ Views on Open Government Data.”(2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

<sup>104</sup>Ibid

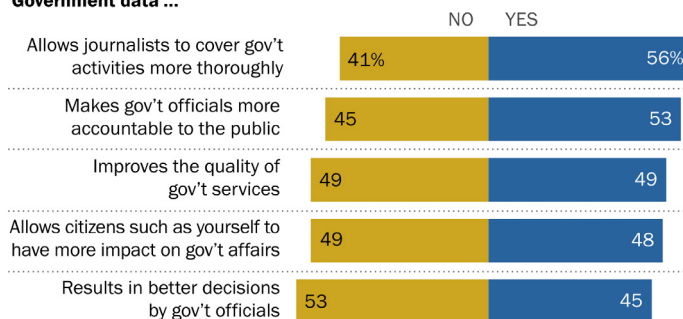
<sup>105</sup>Ibid

Figure 10.

### People Have Mixed Hopes About Whether Open Data Will Improve Things

*% of adults who say these things about the possible impact of government data sharing*

#### Government data ...



Source: Online survey of 3,212 adults in Pew Research's American Trends Panel, Nov. 17-Dec. 15, 2014.

PEW RESEARCH CENTER

From Pew Research Center. "Americans' Views on Open Government Data." (2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

# TAKE ACTION

Cities can ensure that the public recognizes the value of open data by improving the quantity and quality of its open data, and by documenting the positive effects of open data programs.

TITLE	DESCRIPTION
<b>Engage the community about the derived value they desire from open data.</b>	<ul style="list-style-type: none"><li>• Analyze historical public records requests to determine what information the public desires. Information with a high number of public records requests is of clear public interest and therefore a good candidate for open data.</li><li>• Provide potential research directions and applications for open datasets: when publishing datasets, include a list of important questions and applications for which this data could be used. Data is only as valuable as the ways people use it, and if people do not know what might be a valuable use of open data the impacts of open data will be limited. Indeed, a recent study of civic engagement for open data found that a “Problem Inventory” was the most highly sought after engagement approach among open data users.</li><li>• Work with civic hackers to help guide the technical development of the open data platform.</li><li>• Partner with community groups to leverage open data efforts around local issues. Even though these groups may not have as much technical capacity, they often stand the most to gain from new datasets and will have many valuable suggestions for how open data can be used.</li></ul>
<b>Make the data accessible with effective formatting and clear metadata. Data will not provide value if it is context-free and uninterpretable.</b>	<ul style="list-style-type: none"><li>• Clearly describe what each dataset represents and how it is used internally.</li><li>• Provide useful column names and interpretable entries (rather than retaining jargon and shorthand).</li><li>• Reduce redundancy by combining datasets related to the same topic. If releasing multiple related datasets, format them as consistently as possible (e.g., same column names) to make them easy to work with.</li></ul>
<b>Celebrate successful uses of open data.</b>	<ul style="list-style-type: none"><li>• Develop an online showcase that highlights effective and innovative open data uses from the community. This will help build support for open data while also sparking ideas for future projects by showing what is possible.</li><li>• Publish regular reports that include recent uses of open data within the community. Share these positive case studies on social media.</li></ul>

---

<sup>106</sup>Angarita and The City of Cambridge, “Amplifying Civic Innovation: Community Engagement Strategies for Open Data Collaborations.”

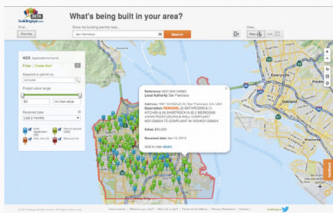
# IN PRACTICE

One of the best ways to increase support for future open data endeavors is to demonstrate that previous efforts have had positive effects on citizens' lives. This helps internal stakeholders recognize the value of their efforts and generates further enthusiasm for open data throughout in the local community.

Two cities that provide such showcases are San Francisco<sup>107</sup> and Philadelphia.<sup>108</sup> Their portals provide links to a variety of applications that have been made using each city's respective open data, showing the diversity of value that open data has created locally.

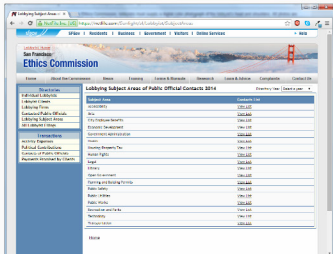
Figure 11a. Screenshot from San Francisco open data showcase

### Featured applications



**Keep an Eye on Construction & Buildings in Your Neighborhood**


Buildingeye.com makes building and planning information easier to find and understand by mapping what's happening in your city. You can sign up to receive alerts about building and construction in your neighborhood.



**Got Lobbying Data?**

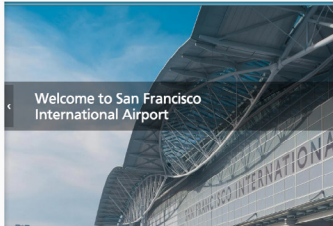
Analyze lobbying activity disclosed by lobbyists registered with the San Francisco Ethics Commission. Learn who pays for lobbyists to influence City policy. Track each City officer with whom the lobbyist made a contact. Discover which candidates are receiving contributions from lobbyists. The datasets are updated every month.

Learn [how to use the API](#) at the Ethics Commission website.



**Explore 311 Data**

The new Open311 Explorer lets you browse requests by type or neighborhood, give it a try!



**SFO API**

The FlySFO.com API grants the web development community access to information about SFO airlines, restaurants, shops and more! Flysfo.com API feeds are web-based and available in .json and XML formats.

City and County of San Francisco, "SF OpenData Showcase," <https://data.sfgov.org/showcase>.


<sup>107</sup>City and County of San Francisco, "SF OpenData Showcase," <https://data.sfgov.org/showcase>

<sup>108</sup>OpenDataPhilly, "Welcome," <https://www.opendataphilly.org/>




Figure 11b. Screenshot from Philadelphia open data showcase

## Featured Projects



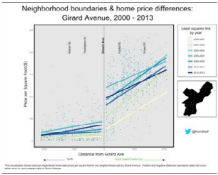
**Greater Philadelphia GeoHistory Network**

provides access to historic geographic resourc...




**Walkshed**

Walkshed is an online application which enable...




**Visualizing Neighborhood Change**

The study of neighborhood change; urban econom...




**School Budget**

Visualize the Philadelphia School District's b...




**Philadelphia Parking Violations**

Newly released data on 4.9 million tickets wri...



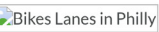
**2012-2014 Crime Incidents (Part I & II)**

Philadelphia Police Department crime incidents...



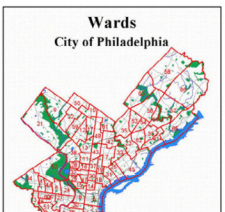
**GoPhillyGo**

GoPhillyGo is a new online mapping tool for t...

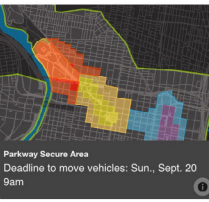


**Bikes Lanes in Philly**

Created using the Bike Network dataset from Op...




**Wards City of Philadelphia**




**Philly Pope Map**

This is an attempt to aggregate the myriad sou...



**Open Budget Explorer**

Budget visualization for the City of Philadelp...



**Philadelphia Blueprint Map**

Cyanotype-inspired map style

OpenDataPhilly, "Welcome," <https://www.opendataphilly.org/>

## 4.2 DEVELOP CONSTITUENCY TRUST BY CONSIDERING PUBLIC EXPECTATIONS AND ACTING AS RESPONSIBLE DATA STEWARDS.

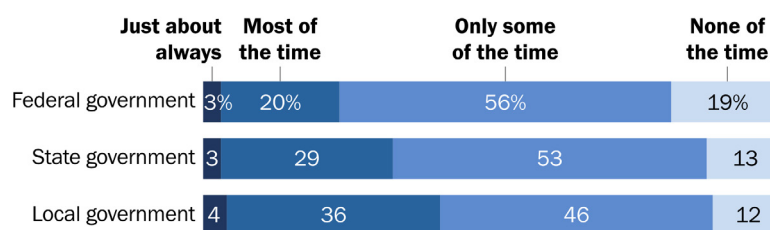
Public trust is critical for supporting open data, yet the public has little confidence in the government to responsibly handle data. Data stewards can earn public trust through effective communication and responsible practices.

As data leaks, cybersecurity hacks, and re-identification attacks grow more common, privacy has become an increasingly tangible and salient issue for the public. Cities face an uphill battle on this public relations front, as local government data initiatives may be conflated with fears related to NSA surveillance. For example, a 2014 Pew Research Center study on public perceptions of privacy found that almost 80% of American adults thought that the public should be concerned about government surveillance.<sup>109</sup> While local governments were not involved in the data collection efforts implicated by the Snowden revelations, much of the public will not distinguish between the branches of government in their perceptions of trust and privacy. Indeed, a 2015 Pew Research Center study found that the public had only marginally more trust in local government than in

Figure 12.

### Majorities Have Low Levels of Trust in Government

% of adults who trust the government ...



Source: Online survey of 3,212 adults in Pew Research's American Trends Panel, Nov. 17-Dec. 15, 2014.

PEW RESEARCH CENTER

From Pew Research Center. "Americans' Views on Open Government Data." (2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

<sup>109</sup>Pew Research Center. "Public Perceptions of Privacy and Security in the Post-Snowden Era." (2014) <http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/>.

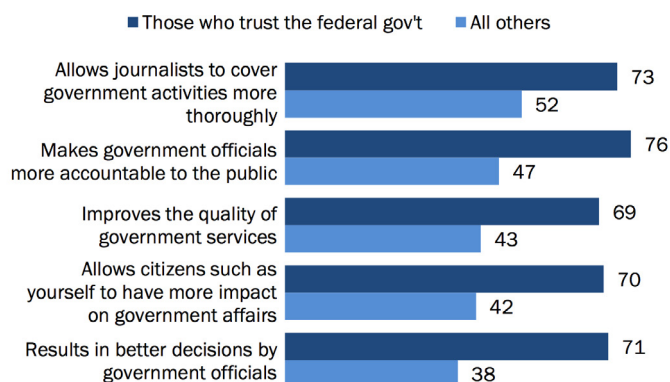
<sup>110</sup>"Americans' Views on Open Government Data."

<sup>111</sup>Susan Gunelius. "How and When UK Consumers Will Share Private Data with Brands." Corporate Eye (2013) <http://www.corporate-eye.com/main/how-and-when-uk-consumers-will-share-private-data-with-brands-infographic/>.

Figure 13.

### Those Who Trust Government Are More Likely to Think There Are Benefits to Opening Government Data

% of adults who believe there are benefits to government sharing data



Source: Online survey of 3,212 adults in Pew Research's American Trends Panel, Nov. 17-Dec. 15, 2014. "Those who trust the federal gov't" refers to those who trust the federal government "just about always" or "most of the time."

PEW RESEARCH CENTER

From Pew Research Center. "Americans' Views on Open Government Data." (2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

state or federal government.<sup>110</sup> Trust in government was critical in driving support for open data, however: individuals with more trust in government also had far greater belief in the benefits of open data.

A pronounced public engagement strategy thus appears critical for cities wishing to re-establish a positive narrative about government data. Indeed, a study of consumer-corporate relationships in the UK shows that the two most important factors determining the public's willingness to share data with certain brands is the perceived trustworthiness of the brand and whether it has a clear privacy policy.<sup>111</sup> Cities are now also in a position to establish a data stewardship relationship with their constituents, and should consider how best to communicate their commitment to security, privacy, and data integrity.

The Pew surveys also provide insights regarding what information the public is comfortable being shared. Over 80% of respondents are comfortable with the government sharing health and safety data on restaurants, but only 22% are comfortable with online disclosure of individual homeowner mortgages.<sup>112</sup> This suggests that citizens' privacy concerns are manifested most acutely when it comes to open data about individuals.

<sup>112</sup>Pew Research Center, "Americans' Views on Open Government Data."

<sup>113</sup>Helen Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life* (Stanford University Press, 2009); "A Contextual Approach to Privacy Online," *Daedalus* 140, no. 4 (2011).

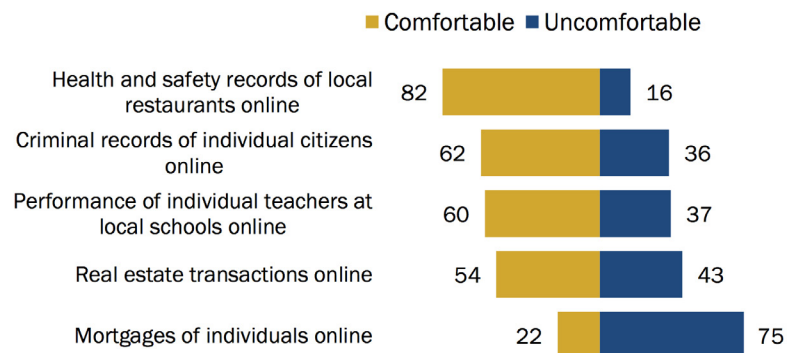
<sup>114</sup>Kirsten E. Martin and Helen Nissenbaum, "Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables," *Columbia Science and Technology Law Review* (forthcoming) (2016).

<sup>115</sup>Helen Nissenbaum, "Privacy as Contextual Integrity," *Washington Law Review* 79, no. 1 (2004).

Figure 14.

## People are Generally Comfortable with Local Government Data Sharing – Until it Hits Close to Home

% of adults who are comfortable/uncomfortable with local government data sharing about these issues



Source: Online survey of 3,212 adults in Pew Research's American Trends Panel, Nov. 17-Dec. 15, 2014.

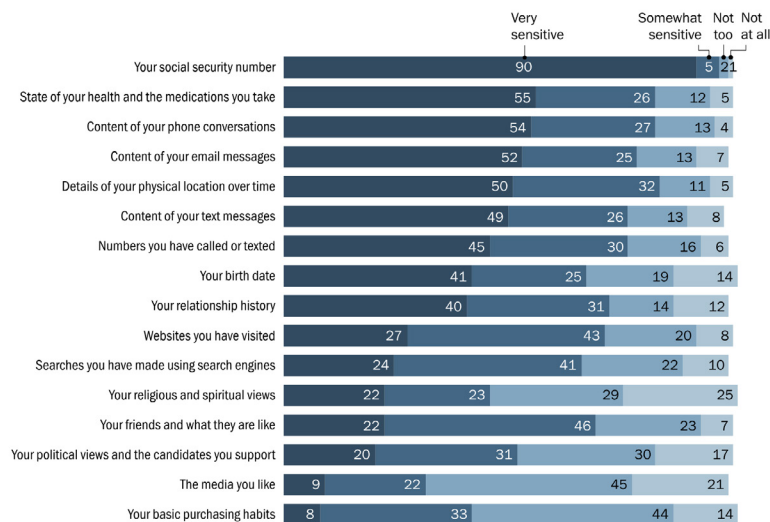
PEW RESEARCH CENTER

From Pew Research Center. "Americans' Views on Open Government Data." (2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

Figure 15.

## Social security numbers, health info and phone conversations among the most sensitive data

% of adults who report varying levels of sensitivity about the following kinds of info



Source: Pew Research Privacy Panel Survey, January 2014. N=607 adults, ages 18 and older.

PEW RESEARCH CENTER

From Pew Research Center. "Public Perceptions of Privacy and Security in the Post-Snowden Era." (2014) <http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/>.

For additional insights regarding public comfort with data releases, data stewards can turn to the work of Helen Nissenbaum for an analysis of privacy concerns in social contexts.<sup>113</sup> Nissenbaum emphasizes how social norms drive our expectations about what information should be available, and who has access to that information.<sup>114</sup> As a recent report explains, “People care about and value privacy – privacy defined as respecting the appropriate norms of information flow for a given context [...] privacy expectations [are] highly dependent on the contextual factors—such as actors receiving the information as well as uses of information. In fact, how the information is used is more important to meeting/violating privacy expectations than the type and sensitivity level of given information.” When data is used for purposes beyond the reason it was originally collected, people are concerned because the social context of that data has changed. Nissenbaum’s approach also explores the limitations of traditional “notice and consent” models, showing that citizens often do not understand the implications of data release disclosures or privacy policies. According to Nissenbaum, transparency as to the fact that data will be released is not enough. Instead, citizens desire and should be empowered with “the right to control information about oneself.”<sup>115</sup> When city open-data initiatives release information that might implicate individuals (such taxi records or location data about crimes), they may violate constituent expectations about control over personal data trails.

# TAKE ACTION

When preparing to release public data that may affect citizen privacy, cities should begin with an understanding of public opinion: awareness of these opinions can guide data processing decisions; prompt informed engagement with citizens before, during, and after release; and improve the means by which released data are branded and communicated to the public. Furthermore, when navigating the ambiguity of what data qualifies as sensitive, cities should incorporate traditional models of privacy preservation (i.e., avoid releasing directly personally-identifiable information) as well as more context-specific social norms that shape public expectations about data.

The three-step process below can help cities manage these tricky issues and determine how best to factor public perceptions into their decisions of when and how to release data.

**1. Determine benefits** — Public support for releasing data (public records requests, etc.) —

	<b>LOW:</b> The public has not expressed any interest in this data or related data.	<b>MEDIUM:</b> The public has expressed mild interest in this data or related data.	<b>HIGH:</b> The public has expressed high interest in this data or related data.
Tangible benefits of releasing data	<b>LOW:</b> There are no clear benefits to releasing this data	Low benefit	
	<b>MEDIUM:</b> There are mild benefits to releasing this data		Medium benefit
	<b>HIGH:</b> There are significant benefits to releasing this data.		High benefit

**2. Determine risks** — Social norms violations of releasing data (i.e., public expectation of privacy) —

	<b>LOW:</b> The data reveals information that is typically made public.	<b>MEDIUM:</b> The data might reveal mildly sensitive information that some members of the public would not expect to be made available.	<b>HIGH:</b> The data reveals sensitive information that the public would not expect to be made available.
Re-identification risks and harms of releasing data	<b>LOW:</b> Re-identification is not likely, nor would it generate any harm for those affected.	Low risk	
	<b>MEDIUM:</b> Re-identification is somewhat likely to occur, and could generate moderate harm for those affected.		Medium risk
	<b>HIGH:</b> Re-identification is likely to occur, and could generate significant harm for those affected.		High risk

### 3. Balance benefits and risks

Benefit

Risk		LOW	MEDIUM	HIGH
	LOW			Release data
	MEDIUM		Possibly release data; weigh the case-specific risks and benefits	
	HIGH	Do not release data		

# IN PRACTICE

InBloom was an education data mining company whose mission was to gather data from schools so that it could study student academic performance and develop individualized curricula. The company quickly raised \$100 million in seed funding after being founded in 2011 and partnered with multiple states. Despite the potential value of this data to improve help schools and teachers, however, InBloom faced organized pushback from groups of parents who were concerned about how their children's data would be used. Parents were wary that the data would not be properly protected from hackers or that it would be sold to data brokers.

Rather than demonstrating to the public that it took these concerns seriously and working to build public trust, "InBloom and the New York State Education Department were arrogant [...] and insensitive to parents who were concerned about their children's data being collected," according to the NYCLU.<sup>116</sup> Despite repeated requests, the company did not let parents opt their children out of the program, instead dismissing all of their fears as misconceptions

Ultimately, these concerns and issues led to the dissolution of InBloom. By the time InBloom announced in 2014 that it would shut down, six of its nine state partners had backed out of relationships with the company.

The downfall of InBloom due to privacy concerns proves the importance of having an effectively communicated privacy program. Because it did not properly consider or respond to public expectations, InBloom faced a significant backlash that ultimately prevented it from achieving its goals. While InBloom felt that the demands for greater privacy would have hindered its progress, in fact a more responsible approach to privacy would have enabled innovation, not prevented it entirely.

This story also highlights other important lessons for data stewards, including:

- People desire agency over how data about them is used. InBloom exacerbated tensions by not allowing privacy-concerned parents to opt their children out of the program.
- Data about certain populations is particularly sensitive. While it is common practice for Internet companies to mine user data and share it with third parties, parents are particularly sensitive about privacy when it comes to their children. Treating student data the same as other information generated significant pushback.
- When data is used for purposes unrelated to the reasons it was collected, people from whom that data was gathered may feel exploited. A 2014 White House report on Big Data states, "As students begin to share information with educational institutions, they expect that they are doing so in order to develop knowledge and skills, not to have their data used to build extensive profiles about their strengths and weaknesses that could be used to their disadvantage in later years."

---

<sup>116</sup>Ariel Bogle, "What the Failure of inBloom Means for the Student-Data Industry," Slate, April 24, 2014. [http://www.slate.com/blogs/future\\_tense/2014/04/24/what\\_the\\_failure\\_of\\_inbloom\\_means\\_for\\_the\\_student\\_data\\_industry.html](http://www.slate.com/blogs/future_tense/2014/04/24/what_the_failure_of_inbloom_means_for_the_student_data_industry.html).

<sup>117</sup>Executive Office of the President. "Big Data: Seizing Opportunities, Preserving Values." (2014) [https://www.whitehouse.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf).



## 4.3 BAKE PUBLIC INPUT INTO ALL ASPECTS OF THE OPEN DATA PROGRAM.

Transparency and responsiveness are critical to building public support and achieving the aims of open data. Incorporating public desires into open data decisions will help ensure that the public understands and supports the program.

Citizens want agency, control, and transparency as to how their data is collected, and how it is used. A May 2015 Pew Research Center study about Americans' views on data collection shows that 93% of adults value "being in control of who can get information about them."<sup>118</sup> Additionally, 90% care about controlling the content of what is collected about them.

Meanwhile, citizens also have high demand for certain information. Data for tracking sex-offenders<sup>119</sup> and city budgets<sup>120</sup> are of constant interest, and the public is particularly concerned when such data is even temporarily withdrawn or obfuscated. Citizens also care about health and safety information about natural resources, restaurants, hospitals, and other city-regulated public spaces.<sup>121,122</sup> Providing open data about these topics will garner public support for open data. Cities are less likely to receive pushback about releasing sensitive information when the data itself is highly sought after or can clearly contribute to broadly desired goals (e.g., improving public safety).

---

<sup>118</sup>Pew Research Center. "Americans' Views About Data Collection and Security." (2015) <http://www.pewinternet.org/2015/05/20/americans-views-about-data-collection-and-security/>.

<sup>119</sup>Milton J. Valencia, "State begins review of 500 sex offender cases," The Boston Globe, February 25, 2016. <https://www.bostonglobe.com/metro/2016/02/24/new-hearings-for-sex-offenders-begin-this-week-could-take-years-complete/dfp465dWlnMKZDGdVRUPnJ/story.html>.

<sup>120</sup>Pamela Martineau, "Open Data Evolution: From Increasing Transparency to Engaging Citizens," Government Technology, March 10, 2015. <http://www.govtech.com/data/Open-Data-Evolution-From-Increasing-Transparency-to-Engaging-Citizens.html>.

<sup>121</sup>Anna Maria Barry-Jester, "What Went Wrong In Flint," FiveThirtyEight, January 26, 2016. <http://fivethirtyeight.com/features/what-went-wrong-in-flint-water-crisis-michigan/>.

<sup>122</sup>Pew Research Center, "Americans' Views on Open Government Data."

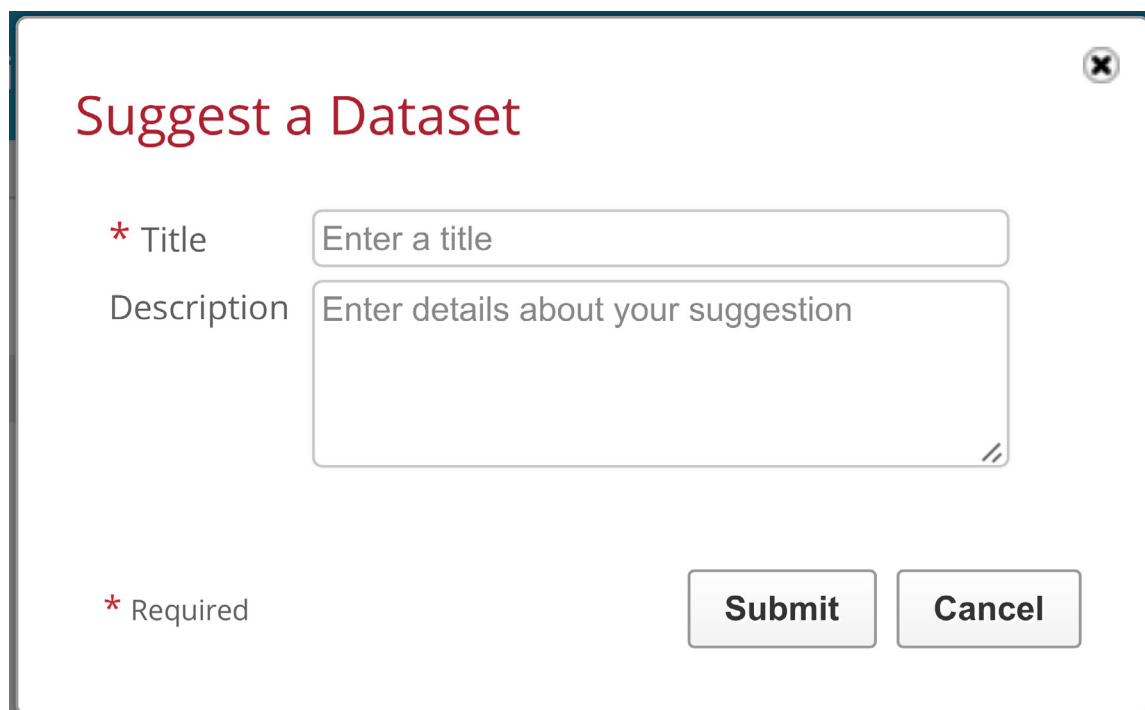
# TAKE ACTION

There are multiple steps that cities can take to incorporate public input into their open data decisions:

- Form privacy committees with members of the public. These groups allow members of the public to have a direct voice in how open data programs approach privacy and ensure that public concerns are accounted for when making open data decisions. These groups should draw upon Institutional Review Boards (IRBs, committees that oversee human subject studies to ensure they are conducted legally and ethically) to develop structures and responsibilities for how to monitor data collection and sharing practices.
- Incorporate public responses on what data is released. This will help ensure that open data programs are responsive to and build value for the public.
- Provide tools for public feedback on open data portals. These could include requests for new data and public comments with questions or concerns about particular datasets.
- Be responsive to public concerns when they arise. Addressing issues openly and honestly will enable a dialogue that will help both sides understand one another's perspective and goals.

Following these strategies will help cities explain future decisions regarding how data is released. Even if re-identification occurs, cities can respond by noting that their data release process was based on priorities that were developed in collaboration with the public.

*Figure 16. Seattle dataset suggestion form*



The screenshot shows a web form titled "Suggest a Dataset" in a dark blue header bar. The form has a white background and a dark blue border. It contains two input fields: a text field for "Title" and a larger text area for "Description". Both fields have placeholder text. At the bottom, there are two buttons: "Submit" and "Cancel". A red asterisk and the word "Required" are located at the bottom left of the form.

**Suggest a Dataset**

\* Title

Description

\* Required

**Submit** **Cancel**

Screenshot from <https://data.seattle.gov/nominate>.

# IN PRACTICE

In 2014, Seattle convened a Privacy Advisory Committee consisting of local technologists, lawyers, and community representatives that was charged with developing principles that could inform the City's privacy policies. This committee created the City's Privacy Principles and helped structure its broader approach to privacy. Seattle's extensive privacy guidelines, which reflect the priorities of the public, are now among the most thorough in the country.

Similar to Seattle, Chicago has convened a privacy committee that is responsible for overseeing its Array of Things project. The project website states,

“Operating as an external, independent review team, the committee will also be consulted whenever there is a request for a new kind of data to be collected. [...] No data will be monitored without the approval of the privacy and security external oversight committee.”

The Array of Things has also pursued other proactive forms of public engagement and in August 2015 released an extensive report detailing all of the ways it had engaged the City's community about the project. This document describes the steps taken to engage residents, summarizes the public feedback received, and outlines lessons learned for future efforts. Resident feedback showed a desire for greater clarity and more thorough policies related to the people and groups involved, data collection and sharing, and public notice. Meanwhile, the report highlights lessons learned, including “to undergo a wider awareness campaign to inform residents of the who, what, where, when, and why of the project before asking residents to react to that project” and stressed that it is “important to communicate what the sensors can't do” (in response to public concerns about potential privacy violations).

This report makes it clear that those responsible for the Array of Things pursued a high level of public feedback and that they built this input into both their policies as well as their plans for engaging the public even more effectively moving forward.

---

<sup>123</sup>Privacy Advisory Committee, City of Seattle <http://www.seattle.gov/tech/initiatives/privacy/privacy-advisory-committee>

<sup>124</sup>Privacy, City of Seattle <http://www.seattle.gov/tech/initiatives/privacy>

<sup>125</sup>Array of Things, “Array of Things,” <https://arrayofthings.github.io>.

<sup>126</sup>*Ibid.*

<sup>127</sup>“Array of Things Civic Engagement Report,” <https://arrayofthings.github.io/engagement-report.html>.

## 4.4 BE TRANSPARENT AND ACCOUNTABLE REGARDING ALL PRACTICES RELATED TO OPEN DATA.

Releasing data is a means toward transparency and accountability, not the end. Achieving transparency and accountability requires thoroughly baking these aims into all aspects of the open data program and carefully documenting those practices.

Open data initiatives are driven by a desire to increase government transparency and accountability. While opening internal data contributes to these aims, it is possible for the process through which data is released to have negative impacts on transparency and accountability.

This is particularly true with regard to privacy. As described in previous sections within this Chapter, most people are nervous about their individual privacy and have low levels of trust in government. This places open data programs in a precarious position: they must release data, but also must be careful not to violate public desires related to privacy or government use of data. Because releasing datasets involves complex questions without a clear answer, open data practices will elicit varied responses from the public. In particular, any actions that cause negative consequences for individual privacy will likely cause the public to become less trusting of government — even if releasing the data makes government more open. By transparently sharing their practices and holding themselves accountable for the effects of their decisions, open data programs can mitigate some of these trust issues.

The best way to ensure that open data programs contribute to transparency and accountability is to bake these goals into every practice and evaluate impact based on contributions to these aims. Instead of viewing the complex challenges associated with open data as roadblocks hindering their progress, open data leaders should embrace them as opportunities for collaborative decision-making with constituents. What better area to pioneer open government decision-making than a program already designed precisely for that purpose?

# TAKE ACTION

Cities can take the following steps to ensure that all of their data practices are transparent. Underlying the following recommendations are explanations: do not just say what was done, but also explain why. Ensure that all descriptions are accessible to non-experts. The public may not agree with every decision, but by sharing how decisions were reached open data leaders will encourage informed deliberation rather than knee-jerk controversy.

TITLE	DESCRIPTION
<b>Data collection</b>	<ul style="list-style-type: none"><li>• Communicate when and where data is being collected, from what sources, and how it may be disclosed in the future.</li><li>• Clarify how data will be used and for how long it will be stored. Equally important can be clarifying how data will not be used. Public concerns about data are often foreseeable, and in most cases cities have no intentions of taking those actions. Cities can stem the tide of backlash by clearly stating the problematic ways in which data will not be leveraged.</li></ul>
<b>Data release</b>	<ul style="list-style-type: none"><li>• Disclose contextual details about data releases so that the public understands the risks and benefits. This includes information about what the data represents and how it was generated, as well as more data-level information such as the accuracy and precision of the information.</li><li>• Explain what actions to protect privacy were considered and taken when releasing data, and the basis for making those decisions.</li><li>• Be transparent about de-identification methods. If published data has been altered from its original form, be sure to report how and why. Otherwise, the public may misinterpret the information provided or draw faulty conclusions.</li></ul>
<b>Responses to public</b>	<ul style="list-style-type: none"><li>• Be upfront about responsibility if privacy issues arise. Explain the assumptions and decisions that went into a particular data release, and why those may or may not have been misguided. This, of course, requires having a privacy evaluation process that can be confidently described and defended.</li></ul>

# IN PRACTICE

Two stories from the City of Chicago highlight best practices for how cities can be transparent and accountable when sharing open data. In November 2016, when the City released data about more than 100 million taxi trips spanning a period of three years, it also published a post describing the data.<sup>128</sup> Included in this report was a thorough discussion of how the data had been altered to protect privacy and improve quality. Not only did Chicago take steps to protect individual privacy in this new dataset (such as delaying publication and masking locations), it also shared how it had done this so that users would understand the city's privacy processes and the data itself. This step should become common practice among cities, especially for datasets that are complex and involve information about individuals.

The City of Chicago was also particularly open and responsive while developing the policies for its Array of Things.<sup>129</sup> As a first step, the City published draft policies and invited the public to provide comments and feedback. Beyond incorporating the public comments into the final policies, the Array of Things also released a document responding specifically to every comment.<sup>130</sup> Questions and responses considered topics such as the location of sensors, the partners involved, potential privacy risks and mitigations, and future plans.

The following shows an example of the questions and responses related to privacy:

QUESTION	ANSWER
<b>This is a concerning piece of wording and implementation of this proposal. This makes me have to ask about the specific management rules of these images - who has access, how long will they be stored, and how do they get deleted? If these images are never deleted, then the entire PII section of this document is void from a technical perspective. With enough images taken over time, one can find an individual based on their clothing, follow them through each image, and eventually determine where they work and where they live. From there, it's pretty easy to figure out the rest of that person's identity. Blurring out images and license plates is not enough. To me, I think it would be better if a smarter solution could be implemented to where images are not even needed for these metrics (i.e. traffic patterns). I don't know what that solution would be, but I'm more afraid of the potential of future harm to be done with these images more than anything.</b>	The policy document has been updated to clarify that image processing for the street-facing cameras will be done on the nodes themselves, and the images will then be deleted - not saved or transmitted. For calibration of image processing software, a fraction of 1% of images will be randomly saved. This limited set of images will contain no sensitive PII. Some may potentially show faces or license plates, and while these are not considered sensitive PII the project has elected nonetheless to limit access to those images to approved individuals who have signed usage agreements, as outlined in the published privacy policy document.

<sup>128</sup>Digital Chicago, "Chicago Taxi Data Released," <http://digital.cityofchicago.org/index.php/chicago-taxi-data-released/>.

<sup>129</sup>Array of Things, "Array of Things Civic Engagement Report".

<sup>130</sup>"Responses to public feedback," <https://arrayofthings.github.io/policy-responses.html>.

QUESTION	ANSWER
I think information sharing should be limited carefully. No data should be downloaded to individual personal devices. This sounds a lot like big brother. If the data is there somebody will access and use it.	No data with any information about an individual will be published. All data management and access within the project team is governed by signed ethics and privacy agreements. These agreements include restrictions on where the data may be processed, including prohibition from storing on personal devices of any kind.
It is the following section which causes me the most concern: "The Array of Things technology is designed and operated to protect privacy. PII data, such as could be found in images or sounds, will not be made public. For the purposes of instrument calibration, testing, and software enhancement, images and audio files that may contain PII will be periodically processed to improve, develop, and enhance algorithms that could detect and report on conditions such as street flooding, car/bicycle traffic, storm conditions, or poor visibility. Raw calibration data that could contain PII will be stored in a secure facility for processing during the course of the Array of Things project, including for purposes of improving the technology to protect PII. Access to this limited volume of data is restricted to operator employees, contractors and approved scientific partners who need to process the data for instrument design and calibration purposes, and who are subject to strict contractual confidentiality obligations and will be subject to discipline and/or termination if they fail to meet these obligations." Of course the question becomes how does the public verify precisely who has such access to the PII data? Will access parameters be modified over time? Specifically, what assurances can one gain that the Chicago Police Department, NSA, or other agencies will not have access to this data?	<p>The documents have been clarified to differentiate between "non-sensitive PII" such as can be found in the public domain, and "sensitive PII," which can identify an individual. The Array of Things has no capability to access or detect sensitive PII, but can detect visual features that are considered to be "non-sensitive PII" such as faces in the public way or license plate numbers.</p> <p>Although not sensitive PII, the privacy and governance policies nevertheless limit who will have access to data, under what circumstances, and for the limited purpose of research and development. The policies also outline how even this potential non-sensitive PII will be controlled, audited, and protected. One important role of the independent external team (Technical Security and Privacy Group, Section 3.4 of the governance policy document) is to audit the project with respect to compliance to these policies.</p>

By publicly responding to all of these questions, those responsible for the Array of Things demonstrated that they recognize that their success relies on more than just the sensors themselves: it is critical that the public support the program and have a voice in its development. The program's transparency in thoughtfully addressing public feedback plays a crucial role in ensuring that the Array of Things truly benefits all Chicagoans.

## 4.5 BUILD SUPPORT FOR NEW INITIATIVES BEFORE ROLLING THEM OUT.

The benefits from sensors and other forms of cutting-edge data collection will be possible only with public support. Implementing successful initiatives requires proactive engagement to educate constituents and address issues before they arise.

Many new smart cities initiatives have great promise but require new forms of data collection. Given low levels of trust in government (see [Section 4.2](#)), some members of the public will understandably be wary of new forms of government surveillance. Rolling out programs without public discussion is likely to compound these constituents' concern and suspicion. On the other hand, other residents will see the benefit of these programs and be less concerned about privacy. Proactive engagement strategies can harness the energy of excited constituents while mitigating the concerns of skeptics.



# TAKE ACTION

It is critical to advertise upfront the value of data collection and open data with the public. The following process can help cities that are considering new data-driven initiatives.

PROCESS	DESCRIPTION
<b>Develop a preliminary privacy policy. Consider the issues described in <a href="#">Section 2.1</a> to generate this plan, and include the following questions.</b>	<ul style="list-style-type: none"><li>• What data will be collected?</li><li>• How will the data be used?</li><li>• Which of these uses are likely to be supported by the public, and which will be opposed? Reduce or eliminate uses that are likely to elicit strong public opposition.</li><li>• Is all of the data being collected critical to enable the desired uses? Limit collection to only those features necessary for the planned applications.</li></ul>
<b>Publicize planned efforts along with the preliminary privacy policy.</b>	<ul style="list-style-type: none"><li>• Make it clear what data will be collected and how that data will be used.</li><li>• Emphasize the social benefits of these uses.</li><li>• Reduce redundancy by combining datasets related to the same topic. If releasing multiple related datasets, format them as consistently as possible (e.g., same column names) to make them easy to work with.</li></ul>
<b>Engage the public by asking for input regarding this privacy policy, with a particular emphasis on determining what data they want collected and what uses they support.</b>	<ul style="list-style-type: none"><li>• If there are skeptics opposed to data collection, try to understand why they oppose these efforts. There may be particular uses that skeptics fear; by developing the privacy policy to prevent these uses, it may be possible to gain their support.</li></ul>

# IN PRACTICE

In November 2013, backed by \$2.7 million from the Department of Homeland Security, the City of Seattle installed sensors designed to form a mesh network that would help law enforcement communicate during emergencies. The City did not provide the public with any details of the technology or its uses, and when residents noticed the devices they became concerned about the surveillance capabilities of this technology. Local newspaper *The Stranger* published an article explaining that these sensors could detect every wireless device nearby and track someone's movements throughout the city. The story also quoted a Seattle Police Department detective saying that he "is not comfortable answering policy questions when we do not yet have a policy." This further inflamed the public response since it appeared that the City was being cavalier about individual privacy and not taking the necessary precautions to protect against surveillance.

The Seattle Police Department responded to this controversy the following week by announcing that it would deactivate the mesh network, stating "The wireless mesh network will be deactivated until city council approves a draft policy and until there's an opportunity for vigorous public debate." Aruba Networks, the company that developed the technology, publicly addressed the situation, explaining, "the mesh product is not capable of reporting on unassociated devices." The City also released a letter from Aruba explaining how the sensor technology does not track unassociated devices nor will it add this capability in the future. Nonetheless, the public remained suspicious about the potential for surveillance and the program was never reactivated.

This story highlights how a lack of transparency and clarity regarding privacy policies can lead to pushback from concerned residents. Because the City was not proactive in engaging the public about the value of this technology and the steps being taken to protect individual privacy, the worst fears of some residents drove an amplified public conversation. Even when it turned out that the technology could not track individuals as many feared, it was too late to garner support for the program. This negative public response ultimately prevented the Seattle Police Department from activating and using the mesh networking system.

This cautionary tale also highlights how to effectively respond to miscues in order to prevent them from occurring again. In addition to managing the fallout from this incident, Seattle also considered how it could better develop policies and practices to protect privacy in the future. In response, Seattle convened their Privacy Program (see [Section 3.1](#)) and Privacy Advisory Committee (see [Section 4.3](#)), both of which have made substantial positive impacts on Seattle's ability to build effective privacy management into their technology initiatives.

---

<sup>131</sup>Brendan Kiley and Matt Fikse-Verkerk, "You Are a Rogue Device," *The Stranger*, November 6, 2013. <http://www.thestranger.com/seattle/you-are-a-rogue-device/Content?oid=18143845>.

<sup>132</sup>*Ibid.*

<sup>133</sup>"The Seattle Police Department Disables Its Mesh Network (the New Apparatus Capable of Spying on You)," *The Stranger*, November 12, 2013. <http://www.thestranger.com/slog/archives/2013/11/12/the-seattle-police-department-disables-its-mesh-network-the-new-apparatus-capable-of-spying-on-you>.

<sup>134</sup>*Ibid.*

# CONCLUSION

As cities across the country release more open data every day, it becomes increasingly important that they protect the privacy of individuals represented in the data. The privacy risks of open data are diverse, arising from many different types of data. Meanwhile, recent research and events have shown the limits of traditional approaches for protecting privacy that rely on removing a small set of direct identifiers. This report responds to these developments and lays out an array of approaches for cities to protect privacy that span technology, policy, and civic engagement.

Cities should:

1. Conduct risk-benefit analyses to inform the design of open data programs.
2. Consider privacy at each stage of the data lifecycle.
3. Develop operational structures and processes that codify privacy management throughout the open data program.
4. Emphasize public engagement and public priorities as essential aspects of open data programs.

Achieving these goals will require municipalities to incorporate privacy protection as a key goal of its initiatives related to data. Ultimately, cities must embrace their role as stewards of individual data, recognizing that protecting the individuals represented in data is a key responsibility of becoming data-driven.

While this report lays out many key guidelines to manage privacy in open data, further efforts are required to ensure cities have the capacity and tools required for effective data stewardship. Most critically, data scientists and cities must work together to develop specific technical guidelines for how to prepare datasets for release. Given that cities release similar datasets, it may be possible to develop standard approaches for how certain types of information should be managed. Although computer scientists are still developing our understanding of the bounds of data privacy, cities will benefit from explicit instructions for mitigating privacy risks that balance benefit and risk based on our current knowledge of effective approaches. As this report has shown, however, data-level approaches are necessary but not sufficient. Cities must continue to develop operational structures and processes that codify privacy management, and share lessons learned about how to create a strong culture of data stewardship. Of particular importance is the need for cities to develop communications and feedback strategies for public engagement about data privacy. Privacy risks in open data are in many cases matters of public perception, and arise when individuals feel uncomfortable with the types of data being collected or released. Data privacy decisions made in a vacuum will be unsuccessful.

As municipal open data initiatives (and other data-driven programs) grow in scale and complexity, so too must their accompanying data governance. Through the approaches described in this report coupled with best practices that develop as more cities and data scientists take action, it will be possible to achieve new gains through data while also protecting individual privacy.

# REFERENCES

- Aitoro, Jill R.** "Defining privacy protection by acknowledging what it's not." *Federal Times*, March 8, 2016 <http://www.federaltimes.com/story/government/interview/one-one/2016/03/08/defining-privacy-protection-acknowledging-what-s-not/81464556/>.
- Angarita, Jennifer, and The City of Cambridge.** "Amplifying Civic Innovation: Community Engagement Strategies for Open Data Collaborations." (2016) [https://docs.google.com/viewerng/viewer?url=https://data.cambridgema.gov/api/file\\_data/f879b5f3-aa03-4e53-8600-7f5270299a62](https://docs.google.com/viewerng/viewer?url=https://data.cambridgema.gov/api/file_data/f879b5f3-aa03-4e53-8600-7f5270299a62).
- Array of Things.** "Array of Things." <https://arrayofthings.github.io>.
- . "Array of Things Civic Engagement Report." <https://arrayofthings.github.io/engagement-report.html>.
- . "Array of Things Operating Policies." <https://arrayofthings.github.io/final-policies.html>.
- . "Responses to public feedback." <https://arrayofthings.github.io/policy-responses.html>.
- Attorney General v. Assistant Commissioner of the Real Property Department of Boston**, 380 623 (1980).
- Bambauer, Derek E.** "Privacy Versus Security." *Journal of Criminal Law and Criminology* 103, no. 3 (2013): 667.
- Barbaro, Michael, and Tom Zeller Jr.** "A Face Is Exposed for AOL Searcher No. 4417749." *The New York Times*, August 9, 2006. <http://www.nytimes.com/2006/08/09/technology/09aol.html>
- Barry-Jester, Anna Maria.** "What Went Wrong In Flint." *FiveThirtyEight*, January 26, 2016. <http://fivethirtyeight.com/features/what-went-wrong-in-flint-water-crisis-michigan/>.
- Bogle, Ariel.** "What the Failure of inBloom Means for the Student-Data Industry." *Slate*, April 24, 2014. [http://www.slate.com/blogs/future\\_tense/2014/04/24/what\\_the\\_failure\\_of\\_inbloom\\_means\\_for\\_the\\_student\\_data\\_industry.html](http://www.slate.com/blogs/future_tense/2014/04/24/what_the_failure_of_inbloom_means_for_the_student_data_industry.html).
- Boston.gov.** "BOS:311." <https://311.boston.gov>
- Broad, Ellen.** "Closed, shared, open data: what's in a name?" <https://theodi.org/blog/closed-shared-open-data-whats-in-a-name>.
- Burwell, Sylvia M, Steven VanRoekel, Todd Park, and Dominic J Mancini.** "Open Data Policy—Managing Information as an Asset." Executive Office of the President, Office of Management and Budget, 2013. <https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>.
- Cable, Dustin.** "The Racial Dot Map: One Dot Per Person for the Entire United States." <http://demographics.coopercenter.org/DotMap/index.html>.
- Chiel, Ethan.** "Why the D.C. government just publicly posted every D.C. voter's address online." *Fusion*, June 14, 2016. <http://fusion.net/story/314062/washington-dc-board-of-elections-publishes-addresses/>.
- City and County of San Francisco.** "DataSF Guidebook: Data Coordinators Edition." (2016) <https://docs.google.com/document/d/1CJ2uZSYEYcPb6bpqr24kcRCV0zDN-9xYE-o7FA23EMk/>.

———. "SF OpenData Showcase." <https://data.sfgov.org/showcase>

———. "Transportation << San Francisco Data." <http://apps.sfgov.org/showcase/apps-categories/transportation/>.

**City of Boston.** "311, Service Requests." <https://data.cityofboston.gov/City-Services/311-Service-Requests/awu8-dc52>.

**City of Philadelphia.** "Open Budget." <http://www.phila.gov/openbudget/>.

**City of Seattle.** "Open Data Policy." (2016) <http://www.seattle.gov/Documents/Departments/SeattleGovPortals/CityServices/OpenDataPolicyV1.pdf>.

**"Citygram."** <https://www.citygram.org>.

**Cranor, Lorrie.** "Open Police Data Re-identification Risks." <https://www.ftc.gov/news-events/blogs/techftc/2016/04/open-police-data-re-identification-risks>.

**de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel.** "Unique in the Crowd: The privacy bounds of human mobility." *Scientific Reports* 3 (03/25/online 2013): 1376.

**de Montjoye, Yves-Alexandre, Jordi Quoidbach, Florent Robic, and Alex (Sandy) Pentland.** "Predicting personality using novel mobile phone-based metrics." In *Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, 48-55. Washington, DC: Springer, 2013.

**de Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, and Alex (Sandy) Pentland.** "Unique in the shopping mall: On the reidentifiability of credit card metadata." *Science* 347, no. 6221 (2015): 536-39.

**Debelak, Jamela.** "ALPR: The Surveillance Tool You've Probably Never Heard Of." May 20, 2013. <https://aclu-wa.org/blog/alpr-surveillance-tool-you-ve-probably-never-heard>

**Digital Chicago.** "Chicago Taxi Data Released." <http://digital.cityofchicago.org/index.php/chicago-taxi-data-released/>.

**Dwork, Cynthia.** "A firm foundation for private data analysis." *Communications of the ACM* 54, no. 1 (2011): 86-95.

**Eland, Andrew.** "Tackling Urban Mobility with Technology." Google Europe Blog, November 18, 2015. <https://europe.googleblog.com/2015/11/tackling-urban-mobility-with-technology.html>.

**Executive Office of the President.** "Big Data: Seizing Opportunities, Preserving Values." (2014) [https://www.whitehouse.gov/sites/default/files/docs/big\\_data\\_privacy\\_report\\_may\\_1\\_2014.pdf](https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf).

**Finkle, Erica, and DataSF.** "Open Data Release Toolkit." (2016) <https://drive.google.com/file/d/0B0jc1tmJAItcR0RMV01PM2NyNDA/>.

**Future of Privacy Forum.** "Future of Privacy Forum." <https://fpf.org>

**Galvin, William Francis.** "A Guide to the Massachusetts Public Records Law." (2013) <http://www.sec.state.ma.us/pre/prepdf/guide.pdf>.

**Gunelius, Susan.** "How and When UK Consumers Will Share Private Data with Brands." Corporate Eye (2013) <http://www.corporate-eye.com/main/how-and-when-uk-consumers-will-share-private-data-with-brands-infographic/>.

**Headd, Mark.** "In Defense of Transit Apps." <https://civic.io/2014/06/13/in-defense-of-transit-apps/>.

**Information Commissioner's Office.** "Anonymisation: managing data protection risk." (2012).

**James, Becca.** "Stop and frisk in 4 cities: The importance of open police data." <https://sunlightfoundation.com/2015/03/02/stop-and-frisk-in-4-cities-the-importance-of-open-police-data-2/>.

**Kiley, Brendan, and Matt Fikse-Verkerk.** "The Seattle Police Department Disables Its Mesh Network (the New Apparatus Capable of Spying on You)." The Stranger, November 12, 2013. <http://www.thestranger.com/slog/archives/2013/11/12/the-seattle-police-department-disables-its-mesh-network-the-new-apparatus-capable-of-spying-on-you>.

———. "You Are a Rogue Device." The Stranger, November 6, 2013. <http://www.thestranger.com/seattle/you-are-a-rogue-device/Content?oid=18143845>.

**Klarreich, Erica.** "Privacy by the Numbers: A New Approach to Safeguarding Data." Quanta Magazine (2012).

**Langton, Lynn, Marcus Berzofsky, Christopher Krebs, and Hope Smiley-McDonald.** "Victimizations Not Reported To The Police, 2006-2010." Bureau of Justice Statistics (2012) <http://www.bjs.gov/index.cfm?ty=pbdetail&iid=4962>.

**Laperruque, Jake.** <https://twitter.com/JakeLaperruque/status/742464398619512832>.

**Marr, Bernard.** "Why Data Minimization Is An Important Concept In The Age of Big Data." Forbes, March 16, 2016. <http://www.forbes.com/sites/bernardmarr/2016/03/16/why-data-minimization-is-an-important-concept-in-the-age-of-big-data/>

**Martin, Kirsten E., and Helen Nissenbaum.** "Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables." Columbia Science and Technology Law Review (forthcoming) (2016).

**Martineau, Pamela.** "Open Data Evolution: From Increasing Transparency to Engaging Citizens." Government Technology, March 10, 2015. <http://www.govtech.com/data/Open-Data-Evolution-From-Increasing-Transparency-to-Engaging-Citizens.html>.

**Massachusetts Public Records Definition.** Mass. Gen. Laws ch. 4, § 7(26).

**"Mugshots."** <http://mugshots.com/>

**Narayanan, Arvind, and Vitaly Shmatikov.** "Myths and fallacies of "Personally identifiable information"." Communications of the ACM 53, no. 6 (2010): 24-26.

**National Information Standards Organization.** "Understanding Metadata." (2004) <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>.

**National Institute of Standards and Technology.** "Guide for Conducting Risk Assessments." (2012) <http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf>.

**Nemani, Abhi.** "Small (City) Pieces, Loosely Joined." <https://medium.com/@abhinemani/small-city-pieces-loosely-joined-5202fb5a93e3>.

**Nissenbaum, Helen.** "A Contextual Approach to Privacy Online." *Daedalus* 140, no. 4 (2011): 32-48.

———. "Privacy as Contextual Integrity." *Washington Law Review* 79, no. 1 (2004): 119-58.

———. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

**Northcutt, Stephen, Jerry Shenk, Dave Shackleford, Tim Rosenberg, Raul Siles, and Steve Mancini.** "Penetration Testing: Assessing Your Overall Security Before Attackers Do." (2006)

**O'Brien, Daniel Tumminelli, Eric Gordon, and Jessica Baldwin.** "Caring about the community, counteracting disorder: 311 reports of public issues as expressions of territoriality." *Journal of Environmental Psychology* 40 (2014): 320-30.

**O'Brien, Daniel Tumminelli, Robert J Sampson, and Christopher Winship.** "Econometrics in the Age of Big Data: Measuring and Assessing "Broken Windows" Using Large-scale Administrative Records." *Sociological Methodology* 45 (2015): 101-47.

**Obama, Barack.** "Transparency and Open Government." The White House, 2009. [https://www.whitehouse.gov/the\\_press\\_office/TransparencyandOpenGovernment](https://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment).

**Okamoto, Karen.** "What is being done with open government data? An exploratory analysis of public uses of New York City open data." *Webology* 13, no. 1 (2016): 1-12.

**Open Data Institute.** "The Data Spectrum." <http://theodi.org/data-spectrum>.

———. "Open Data Institute." <http://theodi.org>.

**OpenDataPhilly.** "Welcome." <https://www.opendataphilly.org/>

**Pandurangan, Vijay.** "On Taxis and Rainbows." <https://tech.vijayp.ca/of-taxis-and-rainbows-f6bc289679a1>

**Peterson, Andrea.** "Why the names of six people who complained of sexual assault were published online by Dallas police." *The Washington Post*, April 29, 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/>

**Pew Research Center.** "Americans' Views About Data Collection and Security." (2015) <http://www.pewinternet.org/2015/05/20/americans-views-about-data-collection-and-security/>.

———. “Americans’ Views on Open Government Data.” (2015) <http://www.pewinternet.org/2015/04/21/open-government-data/>.

———. “Public Perceptions of Privacy and Security in the Post-Snowden Era.” (2014) <http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/>.

**President’s Council of Advisors on Science and Technology.** “Big Data and Privacy: A Technological Perspective.” (2014) [https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast\\_big\\_data\\_and\\_privacy\\_-\\_may\\_2014.pdf](https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf).

**Privacy Tools for Sharing Research Data.** “DataTags.” <http://datatags.org>.

———. “DataTags-Compliant Repositories.” <http://datatags.org/datatags-compliant>.

———. “Private Data Sharing Interface.” <https://beta.dataverse.org/custom/DifferentialPrivacyPrototype/>.

**Project Open Data.** “Project Open Data Metadata Schema v1.1.” <https://project-open-data.cio.gov/v1.1/schema/-accessLevel>

**Reinvent Albany.** “Listening to FOIL: Using FOIL Logs to Guide the Publication of Open Data.” (2014)

**Rosenthal, Brian M.** “Police cameras busy snapping license plates.” The Seattle Times, August 3, 2013. <http://www.seattletimes.com/seattle-news/police-cameras-busy-snapping-license-plates/>

**Rubinstein, Ira S, and Woodrow Hartzog.** “Anonymization and Risk.” Washington Law Review 91 (2016): 703-60.

**Sadetsky, Greg.** “AOL search data mirrors.” [http://www.gregsadetsky.com/\\_aol-data/](http://www.gregsadetsky.com/_aol-data/)

**Schneier, Bruce.** “Data is a toxic asset, so why not throw it out?” CNN, March 1, 2016. <http://www.cnn.com/2016/03/01/opinions/data-is-a-toxic-asset-opinion-schneier/index.html>.

**Schwartz, Paul M, and Daniel J Solove.** “Reconciling Personal Information in the United States and European Union.” California Law Review 102, no. 4 (2014): 877.

**Segal, David.** “Mugged by a Mug Shot Online.” The New York Times, October 5, 2013. <http://www.nytimes.com/2013/10/06/business/mugged-by-a-mug-shot-online.html>.

**Soltan, Ashkan.** <https://twitter.com/ashk4n/status/742466746079010817>.

**Stepanek, Marcia.** “Weblining: Companies are using your personal data to limit your choices—and force you to pay more for products.” Bloomberg April 3, 2000.

**Stinchcomb, Dave.** “Procedures for Geomasking to Protect Patient Confidentiality.” In ESRI International Health GIS Conference. Washington, D.C., 2004.



**Sweeney, Latanya.** "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 5 (2002): 557-70.

———. "Simple Demographics Often Identify People Uniquely." (2000)

**The City of New York.** "NYC Open Data." <https://nycopendata.socrata.com>

**The National Domestic Violence Hotline.** "Who Will Help Me? Domestic Violence Survivors Speak Out About Law Enforcement Responses." (2015) <http://www.thehotline.org/resources/law-enforcement-responses/>.

**Tockar, Anthony.** "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset." <https://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/>.

**Trotter, J.K.** "Public NYC Taxicab Database Lets You See How Celebrities Tip." *Gawker*, October 23, 2014. <http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546>.

**U.S. Census Bureau Center for Economic Studies.** "OnTheMap." <http://onthemap.ces.census.gov>.

**University of Pennsylvania.** "Putting Differential Privacy to Work." <http://privacy.cis.upenn.edu/index.html>.

**Valencia, Milton J.** "State begins review of 500 sex offender cases." *The Boston Globe*, February 25, 2016. <https://www.bostonglobe.com/metro/2016/02/24/new-hearings-for-sex-offenders-begin-this-week-could-take-years-complete/dfp465dWInMKZDGdVRUPnJ/story.html>.

**Vargas, Claudia.** "City settles gun permit posting suit." *The Philadelphia Inquirer*, July 23, 2014.

**Washington State Office of the Chief Information Officer.** "Securing Information Technology Assets." (2013)

**Whong, Chris.** "FOIing NYC's Taxi Trip Data." [http://chriswhong.com/open-data/foil\\_nyc\\_taxi/](http://chriswhong.com/open-data/foil_nyc_taxi/).

**Wood, Alexandra, and Micah Altman.** Personal communication, 2016.

# APPENDIX. OPEN DATA PRIVACY TOOLKIT

This Appendix synthesizes key elements of the report into an Open Data Privacy Toolkit that cities can use to manage privacy when sharing and releasing data. It contains:

- Summary of recommendations ([Executive Summary](#))
- Background on open data privacy risks and vulnerabilities ([Section 1.2](#))
- Background on open data privacy risk mitigations ([Section 1.3](#))
- Risk-benefit analysis form (Sections [1.1](#), [1.2](#), [1.3](#))
- Open data release checklist ([Section 2.3](#))
- Public perceptions management form ([Section 4.2](#))

# OPEN DATA PRIVACY RECOMMENDATIONS

## **Conduct risk-benefit analyses to inform the design and implementation of open data programs.**

- Determine the desired benefits of releasing each element of open data.
- Recognize the limits of de-identification techniques and evaluate the privacy risks of releasing data.
- Consider a diversity of potential mitigations and choose the one best calibrated to the specific risks and benefits of the data.

## **Consider privacy at each stage of the data lifecycle.**

- Collect: Be mindful of privacy before collecting any data.
- Maintain: Keep track of privacy risks in all data stored and maintained.
- Release: Evaluate datasets for privacy risks and mitigate those risks before releasing data.
- Delete: Where appropriate, retire data stored internally, turn off automatic collection, and remove data shared online to mitigate privacy risks that result from the accumulation of data.

## **Develop operational structures and processes that codify privacy management throughout the open data program.**

- Increase internal awareness of and attention to privacy risks.
- Periodically audit data and processes to ensure privacy standards continue to be upheld.
- Account for the unique risks and opportunities presented by public records laws.
- Develop a portfolio of approaches for releasing and sharing data.

## **Emphasize public engagement and public priorities as essential aspects of open data programs.**

- Garner support for open data by sharing the benefits and successful uses of open data.
- Develop constituency trust by considering public expectations and acting as responsible data stewards.
- Bake public input into all aspects of the open data program.
- Be transparent and accountable regarding all practices related to open data.
- Build support for new initiatives before rolling them out.

# OPEN DATA PRIVACY RISKS

THREAT EVENT	DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Re-identification</b>	Re-identification occurs when individual identities are inferred from data that has been de-identified (i.e., altered to remove individual identity from the data), and new information about those re-identified identities becomes known.	Re-identification involves the ability to learn information about individuals that would not otherwise be known. In many cases this new information can lead to a variety of harms for the re-identified individuals such as embarrassment, shame, identity theft, discrimination, and targeting for crime.	In 2000, Latanya Sweeney showed how de-identified health records could be combined with voting registration records to re-identify the health records of most individuals in the US. This meant that it was possible to identify the individual referenced in many health records that were released under the assumption of anonymity. Scientific American describes a notable example: "William Weld, then the [Massachusetts] governor, assured the public that identifying individual patients in the records would be impossible. Within days, an envelope from a graduate student at the Massachusetts Institute of Technology arrived at Weld's office. It contained the governor's health records."
<b>False re-identification</b>	When data is partially anonymous, individuals are at risk of having sensitive facts incorrectly connected to them through flawed re-identification techniques. This is especially likely to occur when open data is of low quality, and contains incorrect information or is difficult to interpret.	Failed re-identification can be as troubling as successful re-identification. Individuals might have incorrect inferences made about them, which could lead to the same harms listed above for re-identification. These harms might be even more severe for false re-identification, since the outcomes will be based on false information or assumptions.	A release of data pertaining to 2013 taxi trips in New York City allowed journalists to determine where celebrities who had been photographed getting in or out of taxis were going to and coming from, along with the fare and tip paid. Surprisingly, many of these trips contained no recorded tip, leading to reports that certain celebrities were stingy and, in response, defenses from these celebrities' agents. Further analysis of the data revealed that many trips simply have no recorded tip, suggesting that the assumption that some celebrities paid no tip was in fact incorrect and due to issues with data quality.
<b>Profile-building</b>	Many companies and other groups compile information about individuals to build a digital profile of each person's demographics, characteristics, habits, and preferences. Open data might contribute new information to these profiles.	Profiles built on data about individuals can be used to analyze and target information to specific segments of the population, thus facilitating algorithmic discrimination and exclusionary marketing.	It has become common practice for companies to target ads to users based on individual preferences, and, in some cases, treat customers differently based on profiles developed by compiling data about those individuals. Bloomberg calls this practice "Weblining, an Information Age version of that nasty old practice of redlining, where lenders and other businesses mark whole neighborhoods off-limits. Cyberspace doesn't have any real geography, but that's no impediment to Weblining. At its most benign, the practice could limit your choices in products or services, or force you to pay top dollar. In a more pernicious guise, Weblining may permanently close doors to you or your business." Open data can contribute new information that feeds online profiles and allows for potential discrimination.

THREAT EVENT	DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Online discoverability</b>	Information that is available online and accessible from an online search.	When information in open data appears in online search results, it appears to a wide audience who might not otherwise have sought out that information. This is a significant change from the past, in which government records were typically available only to those who visited city hall to access them. Many citizens will be concerned when open data associated with their identity can be discovered through online searches for their name or address. Even if people are comfortable with the data being released on an open data portal, they might assume that the data is accessible only to those who seek it out. Exposing information in open data to online search engines can violate this assumption.	Multiple websites today post arrest records, including mug shots, to the Internet. While this information is public record, traditionally one would have needed to go to a courthouse to obtain it. Now one can find this information, even inadvertently, just by searching the name of someone who is listed by mug shot websites. This is especially damaging, New York Times writes, because "Mug shots are merely artifacts of an arrest, not proof of a conviction, and many people whose images are now on display were never found guilty, or the charges against them were dropped. But these pictures can cause serious reputational damage." The Times cites examples such as an individual who was denied a job due to online mug shots that appeared when a potential employer searched his name. These sites typically require fees up to several hundred dollars to have a mug shot removed, a practice that many have called extortion.
<b>Public backlash</b>	Whenever sensitive information is published as open data, the public is likely to respond by blaming the government entity that released the data and losing faith in that entity to act as responsible data stewards.	Public disapproval of open data releases may result from one of the outcomes described above and suggest that the city is not acting with the best interests of its residents in mind. Furthermore, public disapproval detracts from the viability of an open data program. Without public trust in a city to responsibly share data, open data programs will struggle to gain necessary support for releasing data. More broadly, backlashes due to sensitive data releases undermine the public's trust in government.	In June 2016, Washington, DC published online the City's full voter list, which includes names, addresses, and political affiliations. Many people responded with shock and outrage that DC would publish this information in such a widely available format, tweeting with hashtags like "#open_data_fail" and calling the event "horrific." While a public records law mandated that DC release this information, the event nonetheless made many individuals lose faith in DC as a responsible steward of their information.

# OPEN DATA PRIVACY VULNERABILITIES

VULNERABILITY	DATA DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Direct identifiers</b>	Features within a dataset that, on their own, identify individuals. These features (such as name, address, and Social Security Number) have traditionally been known as personally identifiable information (PII).	Because direct identifiers implicate an individual, all of the data tied to that identifier can be connected to the individual in question.	One dataset commonly released by open data programs is property assessments. Because this information includes each property's owner and address (direct identifiers), most records can be connected to an individual. Any information attached to these records (such as property value, renovation history, and violations) can therefore also be traced back to an individual.
<b>Quasi (a.k.a. indirect) identifiers</b>	Features within a dataset that, in combination with other data, identify individuals. The ability to link features across datasets and learn about individuals is known as the mosaic effect.	Seemingly innocuous data can become revealing when combined with other datasets. Because quasi identifiers provide some information about individuals (although not enough by themselves to identify someone), they often facilitate linkage attacks (using the mosaic effect) that combine auxiliary information with quasi identifiers to identify individuals.	In a 2000 study, Latanya Sweeney showed how de-identified health records (containing the quasi identifiers birthdate, gender, and zip code about every individual) could be combined with voting registration records (which contain direct identifiers such as names along with the quasi identifiers mentioned above) to re-identify the health records of most individuals in the US.
<b>Metadata (e.g., behavioral records)</b>	As The National Information Standards Organization describes, "Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information." In a database of emails, for example, metadata contains the sender, recipient, and timestamp of emails. While email metadata does not contain the contents of emails, it can reveal patterns about how people correspond. As such, metadata often comprises behavioral records.	While metadata has not traditionally been seen as sensitive, the President's Council of Advisors on Science and Technology (PCAST) writes, "There is no reason to believe that metadata raise fewer privacy concerns than the data they describe." Although individual metadata records may appear anonymous, large sets of metadata describe detailed and unique patterns of behavior that make it possible to identify individuals and learn intimate details about those people. Behaviors of individuals can be discovered based on auxiliary knowledge (such as paparazzi photographs) or analyzing trends in the data (such as regular appearances at specific addresses). Furthermore, the privacy risks related to metadata are particularly troubling because such data can reveal intimate details of a person's life that would never otherwise be known and that the re-identified individual may never expect to be accessible.	Metadata is particularly sensitive when it is longitudinal, i.e., when multiple records of the same individual can be connected. In a 2015 study of de-identified credit card metadata, computer scientists showed that many people could be uniquely re-identified from records indicating the times and locations of each person's purchases. Because people's movements and spending habits are idiosyncratic and unique, even a small number of records from one person are unlikely to be replicated by anyone else. In particular, the authors found that "knowing four random spatiotemporal points or tuples is enough to uniquely reidentify 90% of the individuals and to uncover all of their records." Another study found that it was possible to predict people's personalities based on their mobile phone metadata.

VULNERABILITY	DATA DESCRIPTION	RISK DESCRIPTION	EXAMPLE
<b>Addresses</b>	Street addresses or location names.	Location data is often highly identifiable and can reveal particularly sensitive details about individuals. Because addresses identify where someone lives or where an event occurred, they are a rich source of information that make it easy to re-identify or learn intimate information about someone. Locations are also easy to link across datasets, facilitating the mosaic effect.	Many cities publish data about 311 requests, which relate to topics such as street and sidewalk repairs, missed trash pickups, animal waste, and pest complaints. Because a typical entry in a 311 dataset includes the address for which the request is made along with a description of the issue, many requests can be re-identified to determine the requester and information about that person's life.
<b>Geographic coordinates</b>	Coordinates that identify a unique location on a map (i.e., latitude and longitude).	Geographic coordinates present the same vulnerabilities as addresses since they translate into locations. Because geographic coordinates do not by themselves reveal a location, however, they may appear to be less sensitive than the addresses they represent. This is misleading, as it is simple to obtain an address from geographic coordinates through a process known as "reverse geocoding."	Crime data is one of the most heavily sought municipal datasets and, in the case of sexual assault-related incidents, one of the most sensitive. In order to protect the identities of victims when sharing open data, many jurisdictions remove the names and addresses associated with sexual assault incidents. However, such data occasionally includes the geographic coordinates of these incidents. Because it is relatively simple to obtain an address from geographic coordinates, this makes the victims of sexual assault highly identifiable. There are significant consequences if sexual assault victims are re-identified, including undue public scrutiny, violation of state shield laws, and potential chilling effects for future reports of sexual assault and domestic violence.
<b>Unstructured fields</b>	Fields that contain comments, descriptions, or other forms of unstructured text (as opposed to structured fields, in which entries must take one of several predetermined values). Photos can also be considered unstructured fields, as there are often few bounds on what information they may contain.	Freeform text fields are often used in unpredictable ways, meaning that their publication may expose unexpected sensitive information.	In 2012, Philadelphia's Department of Licenses & Inspections published gun permit appeals as part of its open data initiative. These permits included freeform text fields in which applicants explained why they needed the permit, and where some people wrote that they carry large sums of cash at night. As a consequence for publishing 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."
<b>Sensitive subsets</b>	Datasets can provide information about diverse populations or events. Each unique type of person or event represents a subset of the data.	Certain categories of people (such as minors and sexual assault victims) within a dataset may be systematically more sensitive than the rest. Information that might be suitable for release with the majority of data might be highly sensitive when it connects to these sensitive subsets.	In 2016, The Washington Post released a report describing how "the names of six people who complained of sexual assault were published online by Dallas police." While the Dallas Police Department did not release "reports categorized as sexual assaults," some cases involving complaints of sexual assault were classified into categories such as "Class C Assault offenses" and "Injured Person." While it may be appropriate to release names in most cases in these general categories, the subsets related to sexual assault require special protections beyond what is needed for the majority of the data.

# OPEN DATA PRIVACY MITIGATIONS

METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
Removing fields	Deleting fields that contain sensitive information.	Removing the addresses from every record in a dataset of police incidents.	Removing fields effectively removes the risks presented by those fields	This approach nullifies any utility made possible by the fields being removed. So the negative impact on utility is large when removing fields that enable valuable uses, but small for less valuable fields.
Removing records	Deleting records that are particularly sensitive, either because of the type of event represented or because of rare (and hence more easily identifiable) features.	Removing records of sexual assault from a dataset of police incidents.	This is an effective way to protect the privacy of those represented in the removed records.	Because only a subset of records have been removed and the rest remain intact, the data remains viable for analysis. However, the removal of records could skew the results or give a false impression about the underlying data. And any analyses that rely on the removed records will be negatively impacted.
Aggregating data	Summarizing data across the population and releasing a report of those statistics.	Reporting the number of crimes that occurred each month rather than releasing data about individual incidents.	Releasing aggregated data effectively protects privacy, as no raw data entries are released.	This has a severe negative impact on utility, as there is no raw data allowing for insights beyond the statistics presented.



METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
<b>Generalizing data</b>	Reducing the precision of fields in order to make each entry less unique.	Reporting addresses by hundred-block increments, block groups, or census tracts.	The less that data is generalized, the easier it is to re-identify someone. Lower levels of generalization (e.g., block group) provide more opportunities for re-identification than higher levels (e.g., zip code). However, while generalizing data can make re-identification more difficult, research has shown that coarsening data has only limited impact.	The more that data is generalized and is characterized at less granular levels, the less useful it becomes. Lower levels of generalization (e.g., block group) provide more useful information than higher levels (e.g., zip code).
<b>k-anonymity</b>	Generalizing fields such that at least k individuals exhibit each feature within those fields. Different traits will require a different level of generalization, depending on how many other entries exhibit that trait.	For k=5, for example, generalizing dates of crime incidents such that every date shown contains at least five events that occurred. If 5 events occurred in a given hour, then the time of those events would be presented as the hour they occurred; if 5 events occurred in a given day, those events would be attributed to the day with no information about the time of day.	As with generalization, the improvement in privacy protection increases as the level of generalization (in this case, the value of k) increases. However, the efficacy of k-anonymity is limited in high-dimensional datasets (those that contain many fields) and in data that contains outliers or rare cases.	As with generalization, the negative impact on utility increases as the level of generalization (in this case, the value of k) increases
<b>Adding noise (a.k.a. random perturbation)</b>	Adjusting data with randomness to offset its original information.	Offsetting geographic coordinates of crime locations by a random distance and direction (generated from probability distributions).	The level of privacy protection increases as more noise is added to a dataset. The impact of noise depends on the density of the population in question: less dense populations require more noise to protect privacy	As more noise is added to a dataset, the less useful it becomes, since the data presented becomes further removed from the events they represent. Furthermore, noisy data can be hard to communicate to the public and may be seen as misleading or an obfuscation of the truth. <i>We recommend against adding noise, and instead suggest generalizing data.</i>

METHOD	DESCRIPTION	EXAMPLE	PRIVACY IMPACT	UTILITY IMPACT
<b>Creating anonymous identifiers</b>	<p>Replacing attributes with randomly generated codes that have no underlying connection to the attribute they replace. This is done through a correspondence table, in which each unique attribute is paired with a random identifier that will replace that attribute wherever it appears.</p>	<p>In a dataset of taxi trips, replacing each unique license plate with its own unique ID number (e.g., a random number drawn from between 1 and the total number of license plates). Every entry containing a given license plate would have the same ID number.</p>	<p>Anonymous IDs can help protect privacy, assuming that the anonymous IDs are randomly generated and have no systematic connection to the attributes they replace (which would occur for example if the numbers were assigned based on the alphabetical order of license plates or a direct hash of license plates). Note that creating anonymous IDs does not protect against re-identifications or inferences based on analyzing patterns of behavior. Furthermore, having any common identifier across all entries related to a specific individual means that once one entry has been re-identified, all entries for that person have also been re-identified.</p>	<p>This approach should have minimal impacts on utility, since it is still possible to track attributes across records.</p>
<b>Differential privacy</b>	<p>Differential privacy is a formal mathematical definition of privacy that provides a provable guarantee of privacy against a wide range of potential attacks. It is not a single tool, but rather a standard of privacy that 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 to be released and used).</p> <p>Theoretical research on differential privacy is rapidly advancing, and the number of practical tools providing differential privacy is continually growing. For these reasons, differentially private tools are becoming an increasingly promising solution for cities to use in combination with other legal and technological tools for sharing data while protecting individual privacy.</p>	<p>Government agencies and corporations currently use differentially private tools to provide strong privacy protection when sharing statistics. The Census Bureau, for example, currently makes some of its data available using non-interactive differentially private tools. Additional tools for differentially private analysis are under development at research institutions.</p>	<p>Differential privacy provides strong guarantees regarding the exact level of privacy risk available through a dataset. In contrast to traditional de-identification techniques that are often designed to address a narrow class of attacks, systems that adhere to strong formal standards like differential privacy provide protection that is robust to a wide range of potential attacks — including attacks that are unknown at the time of deployment — and do not require the person applying the technique to anticipate particular modes of attack.</p>	<p>Differential privacy minimally alters the underlying data, ensuring that the data retains almost all of its utility even after transformation. This feature distinguishes differentially private tools from traditional de-identification techniques, which often require more blunt alterations.</p> <p>In addition to their robust privacy guarantee, differentially private tools have the benefit of transparency, as it is not necessary to maintain secrecy around a differentially private computation or its parameters. Nonetheless, as with the approach above of adding noise, users of differentially private results may struggle to interpret the data. Furthermore, providing data transformed in this way limits the ability to review specific records and might be seen as antithetical to open data.</p>

# RISK-BENEFIT ANALYSIS FORM

DATA FEATURES (ASSETS AND VULNERABILITIES)	ADVANTAGE EVENTS	ADVANTAGE SOURCES	BENEFIT	THREAT EVENTS	THREAT SOURCES	RISK	RISK-BENEFIT RATIO																																															
<p>What are the rows, columns, entries, or sets of entries that may contribute to benefit or risk?</p>	<p>In what forms is the data feature beneficial? How will it be used?</p>	<p>Who might use the data feature?</p>	<p>What is the overall benefit of the data feature?</p>	<p>In what ways is the data feature risky? How might it be abused?</p>	<p>Who might abuse the data feature?</p>	<p>What is the overall risk of the feature?</p>	<p>What is the overall risk-benefit ratio as determined by the benefit assessment and the risk assessment?</p>																																															
<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregated data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregated data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p>BENEFIT</p> <p>RISK</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>M</td><td>L</td><td>L</td></tr> <tr><td>M</td><td>H</td><td>M</td><td>L</td></tr> <tr><td>H</td><td>H</td><td>H</td><td>M</td></tr> </table>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	M	L	L																																																			
M	H	M	L																																																			
H	H	H	M																																																			
<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p>BENEFIT</p> <p>RISK</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>M</td><td>L</td><td>L</td></tr> <tr><td>M</td><td>H</td><td>M</td><td>L</td></tr> <tr><td>H</td><td>H</td><td>H</td><td>M</td></tr> </table>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	M	L	L																																																			
M	H	M	L																																																			
H	H	H	M																																																			
<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p>BENEFIT</p> <p>RISK</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>M</td><td>L</td><td>L</td></tr> <tr><td>M</td><td>H</td><td>M</td><td>L</td></tr> <tr><td>H</td><td>H</td><td>H</td><td>M</td></tr> </table>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	M	L	L																																																			
M	H	M	L																																																			
H	H	H	M																																																			
<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential benefit of the asset (balancing scale and utility)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p><input type="checkbox"/> Individual records</p> <p><input type="checkbox"/> Aggregate data</p> <p>Potential uses:</p>	<p><input type="checkbox"/> Civic hackers</p> <p><input type="checkbox"/> Community groups</p> <p><input type="checkbox"/> Individuals</p> <p><input type="checkbox"/> Journalists</p> <p><input type="checkbox"/> Researchers</p> <p><input type="checkbox"/> Other</p>	<p>LIKELIHOOD</p> <p>What is the probability that the impact will be realized?</p> <p>IMPACT</p> <p>What is the potential risk of the vulnerability (balancing scale and severity)?</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>L</td><td>L</td><td>M</td></tr> <tr><td>M</td><td>L</td><td>M</td><td>H</td></tr> <tr><td>H</td><td>M</td><td>H</td><td>H</td></tr> </table>		L	M	H	L	L	L	M	M	L	M	H	H	M	H	H	<p>BENEFIT</p> <p>RISK</p> <table border="1"> <tr><td></td><td>L</td><td>M</td><td>H</td></tr> <tr><td>L</td><td>M</td><td>L</td><td>L</td></tr> <tr><td>M</td><td>H</td><td>M</td><td>L</td></tr> <tr><td>H</td><td>H</td><td>H</td><td>M</td></tr> </table>		L	M	H	L	M	L	L	M	H	M	L	H	H	H	M
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	L	L	M																																																			
M	L	M	H																																																			
H	M	H	H																																																			
	L	M	H																																																			
L	M	L	L																																																			
M	H	M	L																																																			
H	H	H	M																																																			

DATA FEATURES (ASSETS AND VULNERABILITIES)

What are the rows, columns, entries, or sets of entries that may contribute to benefit or risk?

MITIGATIONS

What are the potential controls to mitigate risk?

RISK-BENEFIT RATIO AFTER MITIGATION

What is the outcome of the risk-benefit analysis after mitigation?

FINAL OUTCOME

What is the final decision for how to release the data?

☐ Remove fields

☐ Remove records

☐ Aggregate data

☐ Generalize data

☐ Anonymize IDs

☐ Other

Mitigation chosen:

RISK

BENEFIT

	L	M	H
L	M	L	L
M	H	M	L
H	H	H	M

☐ Remove fields

☐ Remove records

☐ Aggregate data

☐ Generalize data

☐ Anonymize IDs

☐ Other

Mitigation chosen:

RISK

BENEFIT

	L	M	H
L	M	L	L
M	H	M	L
H	H	H	M

☐ Remove fields

☐ Remove records

☐ Aggregate data

☐ Generalize data

☐ Anonymize IDs

☐ Other

Mitigation chosen:

RISK

BENEFIT

	L	M	H
L	M	L	L
M	H	M	L
H	H	H	M

☐ Remove fields

☐ Remove records

☐ Aggregate data

☐ Generalize data

☐ Anonymize IDs

☐ Other

Mitigation chosen:

RISK

BENEFIT

	L	M	H
L	M	L	L
M	H	M	L
H	H	H	M

# OPEN DATA RELEASE CHECKLIST

	ATTRIBUTE	RISK DESCRIPTION	EXAMPLE(S)	MITIGATING ACTION(S)
<b>Category 1: Individual identifiers</b>	Does the data contain information and attributes directly tied to an individual?	Many types of information can be used to identify individuals within a dataset. Even if a field does not by itself identify an individual, it can be used in conjunction with other fields to do so.	<ul style="list-style-type: none"> <li>• Name</li> <li>• Sex</li> <li>• Race</li> <li>• Address</li> <li>• Birthdate</li> <li>• Phone number</li> <li>• User ID</li> <li>• License plate</li> </ul>	Reduce the precision of these fields or remove them entirely.
	Does the data contain repeated records of an individual's actions?	Behavioral records, often known as metadata, describe detailed and unique patterns of behavior that make it easy to identify individuals and learn intimate details about that person.	<ul style="list-style-type: none"> <li>• User IDs in records of bikeshare usage</li> <li>• License plates in records of taxi trips</li> </ul>	Remove the fields that provide references to individuals, so that records cannot be connected based on the person. Or provide anonymous identifiers in place of these individual IDs, ensuring that they are randomly generated and there is no systematic connection between the original and anonymized IDs (such as alphabetical order).
<b>Category 2: References to location</b>	Does the dataset contain references to locations?	Location data is often highly identifiable and can reveal particularly sensitive details about individuals.	<ul style="list-style-type: none"> <li>• Addresses of incidents in 911 reports</li> <li>• Pickup and dropoff location of taxi trips</li> </ul>	Remove these fields or reduce the precision (i.e., generalize street address into zip code).
	Does the dataset contain geographic coordinates?	Although not human-interpretable, geographic coordinates can be easily mapped to a street address.	Geographic coordinates for the location of 311 requests	Does the dataset contain references to locations?

	ATTRIBUTE	RISK DESCRIPTION	EXAMPLE(S)	MITIGATING ACTION(S)
<b>Category 3: Sensitive fields and subsets</b>	Does the data contain any unstructured text fields?	Unstructured text fields are often used in unpredictable ways, meaning that their publication may expose unexpected sensitive information.	Permit applications that include the applicant's explanation of why the permit is required.	Remove the unstructured fields entirely or evaluate the entries to check for sensitive information.
	Does the data contain any types of records that are particularly sensitive?	Certain categories of records within a dataset may be systematically more sensitive than the rest.	Sexual assault incidents within a dataset of crime incident reports.	Treat these records with particular care, either by removing the entries entirely or removing/generalizing sensitive fields from these entries.
	Does the data contain information that also appears in other datasets?	Connecting information across multiple datasets may reveal sensitive information that is not contained within any individual dataset. This is known as the mosaic effect.	Demographic information (e.g., age, race, and gender) that appears in multiple datasets.	Remove or reduce the precision of any fields that are present in other public data.

# PUBLIC PERCEPTIONS MANAGEMENT FORM

## 1. Determine benefits

Public support for releasing data (public records requests, etc.)

Tangible benefits of releasing data

	<b>LOW:</b> The public has not expressed any interest in this data or related data.	<b>MEDIUM:</b> The public has expressed mild interest in this data or related data.	<b>HIGH:</b> The public has expressed high interest in this data or related data.
<b>LOW:</b> There are no clear benefits to releasing this data	Low benefit		
<b>MEDIUM:</b> There are mild benefits to releasing this data		Medium benefit	
<b>HIGH:</b> There are significant benefits to releasing this data.			High benefit

## 2. Determine risks

Social norms violations of releasing data (i.e., public expectation of privacy)

Re-identification risks and harms of releasing data

	<b>LOW:</b> The data reveals information that is typically made public.	<b>MEDIUM:</b> The data might reveal mildly sensitive information that some members of the public would not expect to be made available.	<b>HIGH:</b> The data reveals sensitive information that the public would not expect to be made available.
<b>LOW:</b> Re-identification is not likely, nor would it generate any harm for those affected.	Low risk		
<b>MEDIUM:</b> Re-identification is somewhat likely to occur, and could generate moderate harm for those affected.		Medium risk	
<b>HIGH:</b> Re-identification is likely to occur, and could generate significant harm for those affected.			High risk

## 3. Balance benefits and risks

Benefit

Risk

	<b>LOW</b>	<b>MEDIUM</b>	<b>HIGH</b>
<b>LOW</b>			Release data
<b>MEDIUM</b>		Possibly release data; weigh the case-specific risks and benefits	
<b>HIGH</b>	Do not release data		