CENTER FOR DEMOCRACY & TECHNOLOGY

MAKING GOVERNMENT DATA PUBLICLY AVAILABLE

Guidance for Agencies on Releasing Data Responsibly

August 2022

AUTHORS:

Hugh Grant-Chapman
Hannah Quay-de la Vallee

With contributions from Cody Venzke and Elizabeth Laird

The Center for Democracy & Technology (CDT) is a 27-year-old 501(c)3 nonpartisan nonprofit organization that fights to put democracy and human rights at the center of the digital revolution. It works to promote democratic values by shaping technology policy and architecture, with a focus on equity and justice. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

As governments expand their use of technology and data, it is critical that they do so in ways that affirm individual privacy, respect civil rights, foster inclusive participatory systems, promote transparent and accountable oversight, and advance just social structures within the broader community. CDT's Equity in Civic Technology Project furthers these goals by providing balanced advocacy that promotes the responsible use of data and technology while protecting the privacy and civil rights of individuals. We engage with these issues from both technical and policy-minded perspectives, creating solutions-oriented policy resources and actionable technical guidance.

Endnotes in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Government agencies rely on a wide range of data to effectively deliver services to the populations with which they engage. Civic-minded advocates frequently argue that the public benefits of this data can be better harnessed by making it available for public access. Recent years, however, have also seen growing recognition that the public release of government data can carry certain risks. Government agencies hoping to release data publicly should consider those potential risks in deciding which data to make publicly available and how to go about releasing it.

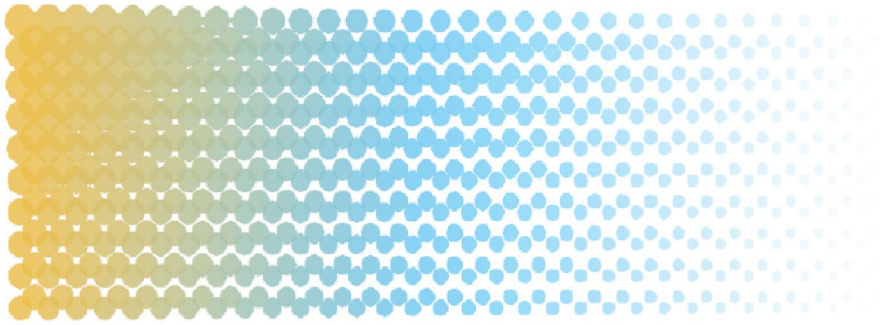
This guidance offers an introduction to making data publicly available while addressing privacy and ethical data use issues. It is intended for administrators at government agencies that deliver services to individuals — especially those at the state and local levels — who are interested in publicly releasing government data. This guidance focuses on challenges that may arise when releasing aggregated data derived from sensitive information, particularly individual-level data.

The report begins by highlighting key benefits and risks of making government data publicly available. Benefits include empowering members of the general public, supporting research on program efficacy, supporting the work of organizations providing adjacent services, reducing agencies' administrative burden, and holding government agencies accountable. Potential risks include breaches of individual privacy; irresponsible uses of the data by third parties; and the possibility that the data is not used at all, resulting in wasted resources.

In light of these benefits and risks, the report presents four recommended actions for publishing government data responsibly:

1. **Establish data governance processes and roles;**
2. **Engage external communities;**
3. **Ensure responsible use and privacy protection; and**
4. **Evaluate resource constraints.**

These key considerations also take into account federal and state laws as well as emerging computational and analytical techniques for protecting privacy when releasing data, such as differential privacy techniques and synthetic data. Each of these techniques involves unique benefits and trade-offs to be considered in context of the goals of a given data release.



INTRODUCTION:
Why is it important to make government data publicly available? Understanding the benefits and risks

Government agencies* routinely collect, store, and analyze data** pertaining to the services they provide and the populations with which they interact. Such information is essential for a wide range of core responsibilities,[1] from coordinating service delivery to meeting community needs,[2] improving operational efficiency,[3] and reacting to changing external contexts.[4]

Community organizers, civil society advocates, and researchers have long advocated increasing the public availability of this data to strengthen democratic governance, institutional accountability, and the ability of communities to solve problems.[5, 6] These calls to action are rooted in the idea that government data is a public good and therefore ought to be made available to the widest audience possible to better reap the benefits of data-driven decision-making.[7] Early efforts to make government-held data publicly available contributed to the passage of freedom of information laws, and more recent advocacy has lent momentum to the open data movement.[8] Such efforts seek to improve the quality of public agency service delivery through accountability and research, while also reducing the need for redundant data collection across organizations with adjacent goals.[9]

The push for publicly available government data, however, has also drawn attention to the risks of careless distribution and use. Chief among these risks is the potential to compromise individuals' privacy by disclosing their personal information.[10] Many government agencies deal

* This report is targeted at state and local government agencies that deliver services to individuals.

** This report defines "government data" as information held by a government agency. Publicly available government data (i.e., government data that has been published or publicly released) refers to government data that is freely and readily accessible to any member of the public.

with sensitive information that could adversely affect community members if made public, which could lead to risks of discrimination,[11] loss of trust,[12] and even threats to personal safety.[13, 14] Moreover, modern reidentification technologies render seemingly innocuous personal data categories vulnerable to linking and exposure of unanticipated inferences that could be exploited by malicious actors.

Because of these factors, a risk–benefit analysis can help inform agency decisions about what data to prioritize making publicly available, which requires an understanding of the benefits and risks associated with public government data.

RELATIONSHIP BETWEEN PUBLIC GOVERNMENT DATA AND OPEN DATA

Many of the recommendations in this report draw from the principles of *open data*. Open data, and open government data in particular, refers to data that has been published in a manner that aligns with open data principles, which themselves are the well-defined products of a longstanding advocacy agenda.[15]

Although this report's use of "publicly available government data" overlaps with the common understanding of "open government data," the concept of open data is distinct and more specific than publicly available data.[16, 17] For instance, to be considered "open data," government data must be machine-readable (see Appendix II on page 30 for more information), "complete" (which refers to disaggregation requirements), and ideally available in an established data repository.[18] These requirements are generally best practices for data that is being published but are not strictly necessary for any release of government data (for instance, privacy considerations may prevent the release of fully disaggregated data). Due to these distinctions, this report uses the more general term, "publicly available government data."

BENEFITS OF MAKING GOVERNMENT DATA PUBLICLY AVAILABLE

Government agencies publish data because it provides them several important benefits and can simultaneously serve the public interest.

Empowering individuals and the general public

Government data can benefit individuals and the general public at large directly. Robust data can empower individuals and organizations to make more effective decisions about how to govern their own lives and improve their communities. The Community Development Department in Madison, WI, worked with the Sunlight Foundation to enable nonprofits to more easily find data sources to help them perform their missions more effectively.[19] Similarly, data can be used by community members themselves to make more effective decisions. These decisions may include determining how to vote on candidates or ballot initiatives or figuring out which city programs or initiatives might best serve their own needs.[20]

Supporting research on program efficacy

Academic research institutions and civil society researchers often rely on government data to conduct civic and social science research, which is especially helpful for understanding community needs and the impacts of policy interventions. Public data access can support these inquiries by helping researchers evaluate service delivery programs over long time frames,[21] compare similar programs from different geographic regions,[22] and understand the impacts of certain external factors on program efficacy.[23] Such research can in turn improve the quality of an agency's services.

Despite these benefits, many research questions require access to more granular data than can be publicly released (for instance, due to privacy concerns).[24] These scenarios likely call for formal data-sharing agreements (discussed more on page 16), but public access to higher-level data can inform initial research questions and spotlight issues for deeper inquiry.

USING PUBLICLY AVAILABLE DATA TO ASSESS THE IMPACT OF A NEW YORK CITY HOUSING SUPPORT PROGRAM

Researchers at Princeton University used publicly available housing court records to assess the impact of New York City's Universal Access (UA) program, which provides legal representation to low-income tenants in housing courts, on court outcomes.[25, 26] The findings were promising: "The UA program increased the likelihood that tenants in their sample had legal representation by 88%. By comparing outcomes within UA ZIP codes, they found eligible tenants who secured legal representation following the introduction of the program were 62% less likely to be subject to possessory judgments and 72% less likely to have eviction warrants issued against them."[27] This kind of research plays a critical role in identifying effective government programs and encouraging civic leaders to expand their use.

Supporting the work of adjacent organizations

Publicly available government data related to a given sector or population can be used by other organizations providing adjacent services to inform and support their work. Government agencies often coordinate with nonprofit partners or other agencies to meet the needs of target audiences.[28] Through such cross-agency or public–private partnerships, adjacent organizations can use government data to help identify unmet needs, provide specialized care, avoid duplicative efforts, and coordinate complementary services. By making this data publicly available, agencies and organizations can share data-driven insights and direct resources to their most efficient uses without requiring a full-fledged data-sharing agreement.[29]

Reducing agencies' administrative burden

Public access to government data can reduce agencies' administrative burden by minimizing ad hoc requests for data. Setting up a public data access system prompts agencies to pre-vet the suitability of their data for public audiences, establish a streamlined process for releasing data, and document which data requires formal data-sharing agreements to access. This added capacity can allow agency staff to focus on deeper insights from data that go beyond descriptive analysis or topline findings. Implementing a system of public data access will not fully eliminate the need for ad hoc requests since external parties may request data that is not available through public-facing systems. Nevertheless, the more data that agencies are able to open up to easy public access, the fewer ad hoc requests they will likely have to address.

FEDERAL AND STATE PUBLIC DATA LAWS

Appendix III on page 31 includes a chart highlighting some examples of laws related to government entities' responsibilities to provide publicly available data and to protect individual privacy. Although the public data laws are diverse, three key trends emerge in these requirements:

1. **Agencies are required to make records and data publicly available.** Many of the laws listed in the chart include public data as a requirement for agencies. Although the respective laws or guidance might provide substantial discretion to agencies to implement the requirements for publicly available data, the overall trend in many of the laws is for agencies' data and records to be public by default. The public availability of data may range from mandatory publication to permitting disclosure upon request.

2. **Publicly available data must adhere to privacy protections and ethical data practices.** Public data laws also require that data be made publicly available in ways that not only protect individual privacy but also meet certain data ethics requirements. For example, the federal Confidential Information Protection and Statistical Efficiency Act limits data use to "statistical purposes" and prohibits other uses such as enforcement actions based on the data it allows to be collected. Other rules may exempt sensitive information from the requirement that data be made publicly available.

3. **Publicly available data is increasingly being used across agencies to provide services.** Agencies are increasingly looking for ways to share data to effectively provide governmental services. This data may be provided in the form of publicly available aggregate data or private individual information; in either case, sharing agencies are often required to adhere to policies regarding use restrictions, privacy protections, and notice to affected individuals.

Although many public data rules reflect these trends, legislative and regulatory efforts demonstrate a variety of approaches. For example, at the federal level, the Privacy Act and the Computer Matching and Privacy Protection Act include legislative requirements for agencies to use and share individual data only in alignment with the purpose for which the data was originally collected. Those requirements are implemented by agencies through formal rulemaking, identifying their respective "systems of records." Similarly, the Foundations for Evidence-Based Policymaking Act is implemented through informal memos issued by the Office of Management and Budget in conjunction with formal solicitations for public input by other executive agencies. States reflect a similar multimodal approach, with some issuing public data requirements through legislation and others detailing those requirements in guidance or even administrative handbooks.

Advancing government transparency

With access to government data, members of the public can improve their understanding of civic systems, including how agencies are run, what kinds of services they provide, and what outcomes they are achieving.[30] This transparency is essential for holding government agencies publicly accountable for providing quality services. Third parties can use public data to evaluate whether agencies are adequately serving their mandated purpose and identify irresponsible practices or inefficiencies within government bureaucracies that would otherwise be opaque. Wasteful or redundant uses of public funding,[31] important but underfunded programs,[32] instances of inequitable service delivery, and biased or discriminatory behavior within agencies[33] can all be potentially identified by analyzing government data.

Data journalists and other reporters can also use publicly available government data to inform their reporting and improve the quality of public debate, particularly around government service providers. Open access to public data is particularly important for journalists since these individuals may not have the same longstanding institutional relationships that academic researchers or organizations involved in public–private partnerships have. Legal rules for public data range from mandating publication and permitting access upon request to prohibiting disclosure of sensitive information. Systems for public data access can help to level the playing field for journalists by providing a consistent, uniform process for a diverse range of individuals and organizations to analyze government data, rather than forcing them to rely on interpersonal or institutional relationships.

RISKS AND COSTS OF MAKING GOVERNMENT DATA PUBLICLY AVAILABLE

Despite the benefits described in the previous section, providing public access to government data does not come without risks. As such, government administrators considering whether to release data publicly should carefully evaluate these possibilities to minimize potential negative consequences.

Breaches of individual privacy

One of the most commonly cited potential harms of public data sharing is breach of individual privacy. Government agencies frequently collect, hold, and use individual-level information. Data that can be directly linked or used to identify a specific individual, referred to as personally identifiable information, is protected by that individual's right to privacy.[34] When personal information is publicly disclosed, individuals can face financial loss,[35] reputational harm,[36] discrimination or inequitable treatment,[37] emotional distress,[38] and threats to physical safety.[39]

The risks of privacy breaches are magnified when the data to be shared publicly includes data that could jeopardize individuals' safety, security, or well-being. Sensitive data categories include data that could lead to discrimination, exploitation, or unjust treatment if disclosed publicly — data related to sexual orientation, immigration status, residential status, and past interactions with law enforcement are a few such examples.[40] Some data, such as data related to victims of domestic abuse or current residential address, is sensitive because it may give rise to safety risks.[41]

Another concern with publicly releasing data is the potential for loss of community trust. This loss of trust has tangible impacts: Concerns about privacy protection may discourage individuals from seeking out public services,[42] particularly when institutions providing these services have poor track records of responsibly using community data. These negative impacts often play out, for instance, when unreleased data is breached.[43] For all of these reasons, privacy protection should be a top priority for any public agency considering releasing data (specific privacy protection strategies for publicly released data are discussed later in this report).

REIDENTIFICATION RISKS

In the past, government agencies seeking to publish their data have attempted to address privacy concerns by removing key elements of datasets to ostensibly deidentify the data and prevent it from being associated with specific individuals. However, data science research has increasingly demonstrated how data-linking technologies can reidentify seemingly anonymous data.[44, 45, 46, 47] Broadly, the efficacy of privacy-protecting techniques depends on the specifics of the dataset and categories of data in question — some kinds of data (such as location data) are essentially impossible to successfully deidentify.[48, 49] As government agencies consider publishing individual data, they should ensure that any anonymization process they apply meets rigorous criteria and is robust against reidentification efforts (discussed more on page 19).

Irresponsible interpretation and use practices

Breaches of individual privacy are not the only risk associated with public access to government data. Published data can be fully privacy-protective and still be used irresponsibly by third parties in ways that jeopardize the public good.

One such risk is that important contextual information will be removed, distorted, or ignored during the process of third parties accessing and analyzing the data. A similar risk exists when sharing data of poor or dubious quality or data that reflects systemic biases. Data that is analyzed without consideration of contextual information increases the likelihood of misinterpretation, which can produce inefficient or harmful outcomes if those faulty inferences are used to inform policy decisions (see text box on page 22). To be clear, these concerns can be addressed and should not accepted as a reason to not release data publicly, but they do point to the need for robust documentation: When releasing data, agencies should clearly publish accompanying information about the limitations of the dataset and the kinds of analysis for which it is, and is not, suitable.

Many data sources and datasets are also vulnerable to misuse because they reflect systemic inequities or could be used to reinforce systemic biases. Data that describes social systems and patterns inevitably reflects the power structures and other biases inherent in the systems it represents. To note one example, students from historically marginalized backgrounds have long been over-represented in data on K–12 school disciplinary practices; these kinds of datasets should thus be treated sensitively because they reflect trends of bias in disciplinary decision-making.[50] It is important that these kinds of discriminatory practices be exposed to the public by publishing data, which can help drive public attention to addressing these inequities. At the same time, other secondary uses should be approached with caution to avoid perpetuating these same biases in subsequent data-driven decision-making, especially in algorithmic systems.

EXAMPLE OF IRRESPONSIBLE USE RISK

In 2010 and 2011, the *L.A. Times* sparked controversy by publicly releasing evaluations of 6,000 Los Angeles Unified School District (LAUSD) elementary school teachers by name based on their students' performance on standardized tests.[51, 52] The published data, which resulted in public shaming and harassment of low-performing teachers, was widely criticized for methodological flaws and lack of attention to contextual information.[53] Moreover, subsequent research on the impact of this decision suggested that it actually exacerbated educational inequities within LAUSD.[54] These harms demonstrate the importance of evaluating the risk of irresponsible use when publishing data.

Financial and labor costs in preparing and releasing government data

Making government data available for public access can be a resource-intensive process, depending on the kind of data to be published and the agency's existing data governance infrastructure.[55] These costs create the risk that agencies will invest time and money — which could otherwise be used for different goals — into preparing and releasing data that then is underused or not used at all. As this report discusses on page 15, publishing data with a clear purpose in mind minimizes the risk that data goes unused upon release and should be factored into an agency's cost–benefit analysis when deciding which data to prioritize for publishing, along with the resources needed to publish data (see page 23).

RECOMMENDED ACTIONS FOR RESPONSIBLY PUBLISHING GOVERNMENT DATA

To increase the benefits of publishing data and minimize potential risks, government agencies should incorporate four key actions into their data release strategy:

1. **Establish data governance processes and roles;**
2. **Engage external communities;**
3. **Ensure responsible use and privacy protection; and**
4. **Evaluate resource constraints.**

These actions span all stages of the data publishing process, from initial decisions on whether to release data through final launch and maintenance.

ESTABLISH DATA GOVERNANCE PROCESSES AND ROLES

Government agencies preparing to release data should establish data governance processes[56] and roles[57, 58] for handling data, which can improve decision-making while reducing the risk of irresponsible data use.[59]

Key steps include:

1. **Determine the decision-making process:** An agency's data governance system should specify the steps of the process to release data publicly and roles for each step.[60] Decisions, and related roles, can include determining which datasets to prioritize for public release, engaging with external communities, processing and reviewing data, and developing an effective publication strategy.

2. **Choose which data to prioritize for public release:** A given agency likely houses many different kinds of data that could potentially be prepared for release, but since publishing data requires time and resources, the agency will need to prioritize bodies of data that add the most public value.[61] An important step in evaluating these prioritization decisions is to consider at the outset the purpose(s) for which the data will be published. Proposed data releases can then be evaluated through a cost–benefit analysis:[62] Do the proposed purposes for the data to be published justify potential privacy risks, resource requirements, and other considerations involved in releasing it, and are there ways these factors can be minimized?

3. **Analyze data to prepare it for release:** Most likely, the data to be released will require processing before it is fit for publishing. This step can involve cleaning and structuring the data,[63] applying manipulations to protect individual privacy,[64] and documenting metadata and contextual information.[65] Once this processing is complete, trained agency staff should thoroughly review the data to ensure robust protection of individual privacy.

4. **Implement a publishing and outreach strategy:** Agencies should choose how they will release the data based on their goals and the unique trade-offs of possible release formats. These can include application programming interface (API) access, bulk dataset download, data visualizations or dashboards, or a combination of approaches (discussed on page 17).[66] An agency's release strategy should also include a plan for promoting the data's availability and attracting public interest.

5. **Update and maintain the data after it has been released:** Data that has been released may need to be updated over time, and the agency in question should establish an updating process at the outset. The agency should design this updating process in a way that clearly communicates when a dataset has been most recently updated and what has changed.

CREATING AND COMMUNICATING TIERS OF DATA ACCESS

Government agencies can integrate publicly available data into a tiered model of data access alongside other categories of data with more limited access.[67] Although this report focuses on the first tier of access (public, proactively released data), understanding tiered systems overall can help inform decisions about how, and why, data is released.

A framework for access tiers can quickly convey how widely accessible a given dataset is and why it is categorized as such. The tier to which given data is assigned is determined by characteristics of that data (e.g., how much private information is disclosed, whether it includes sensitive information categories). Tiers could include:

1. **Public, proactively released data:** Data in this category is available to the widest range of audiences. It includes data that is accessible to anyone directly through the agency website or another public mechanism for any use case. Because this data can be viewed and used by anyone, it requires thorough vetting to ensure that no private information is disclosed.

2. **Data available publicly upon request:** Data in this category is available to any member of the public but is distributed only upon request (rather than proactively). Request criteria can vary in complexity, but data in this category is eligible to be released so long as a request has been filed — no review process is necessary.

   One example of this demarcation can be found in the District of Columbia's Data Policy, which delineates data that can be shared but must be requested because "publication of the dataset on the public Internet and exposure to search engines" could raise concerns such as "jeopardiz[ing] the safety, privacy, or security" of people identified in the dataset; "impos[ing] an undue financial or administrative burden on the agency"; or posing legal liability issues.[68]

3. **Data shared with select institutions or individuals:** Data in this category is not suitable for public access, but agencies may consider sharing it privately with select institutions or individuals by implementing an approval process or requiring a data-sharing agreement. Because this information is more sensitive, data-sharing agreements may be conditional upon application review, which could involve guarantees of privacy and security standards at institutions that will receive the data. Alternatively, data in this tier could be distinguished by restricted access methods such as trusted execution environments, which allow researchers to analyze data and run code in a restricted, secured context that limits the risk that data will be leaked or breached.[69] However agencies choose to share this category of data, it is important that they vet partners before sharing the data. Vetting should consider elements such as whether partners possess the expertise to conduct effective research, the technical resources needed to keep the data private, whether the research will benefit the communities whose data is shared, etc.

4. **Data that is not shared at all:** Data in this category is most sensitive and is not shared outside of the agency. Access to this data may be restricted to only agency staff members with certain clearance.

Some states have laws that establish tiers of access such as those listed previously, as well as criteria to use for sorting agency data into tiers.[70] In these cases, agencies should comply with state law and make use of existing frameworks. For agencies that lack an established tiered framework for data access, administrators should consider establishing their own system to inform and document how their data can be accessed.

ENGAGE EXTERNAL COMMUNITIES

Robust, proactive external engagement safeguards against irresponsible approaches to releasing government data. External audiences should be consulted early in the release planning process to inform what data to release and how to release it (including whether to release the data at all). This engagement should continue through the process of data preparation, publication, outreach, and maintenance to increase visibility of the data and tailor its release format for external audiences' anticipated use. The Center for Democracy & Technology's guidance on community engagement and data sharing includes a comprehensive discussion of how to design and conduct an effective engagement strategy.[71]

Engage communities represented in the data

Agencies considering releasing data should prioritize engaging the populations that the data describes. Because the government data in question was initially collected from these individuals, they may be affected by future decisions that are based upon that data. Moreover, when these individuals originally submitted their data and potentially consented to its use, they may not have anticipated secondary use cases that arise from publishing this data.[72] As a result, government agencies should proactively consult affected communities about whether their data should be made public and in what form. In practice, engagement strategies should be responsive to the specifics of the community being consulted: Datasets that describe nationwide populations will require different kinds of engagement than local-level datasets, and datasets involving vulnerable populations or sensitive data categories may require additional engagement considerations.[73]

It is worth noting that soliciting individuals' consent for an expansive or open-ended range of potential uses of data when it is first collected is not an adequate strategy for evaluating community support. Rather than treating community consent as a check-the-box exercise, government administrators should work together with communities to determine specific goals and methods around how their data will be used. This proactive engagement also helps identify which applications of community data are most valuable to the community.

Engage public audiences that will use the data

External stakeholders — civil society organizations, academic researchers, community advocates, journalists, and general public audiences, to name a few — are the audiences that will make use of the data upon release. As such, government agencies should proactively engage these audiences to determine what kind of data they are most interested in receiving, how they plan to use it, and which formats and contextual information would make it most useful.[74, 75] Once the data is ready to be released, agencies should design a release strategy with the target audiences in mind. Understanding which audiences are requesting data, their needs and intended uses of the data, and their existing technological capacity will help agencies publish their data in a way that maximizes its potential for beneficial use.

WHO USES PUBLIC GOVERNMENT DATA?

Movements to increase access to public government data often focus on using it to support research to make informed policy and practice decisions. The traditional perception has been that researchers are professors and others housed in universities and other academic environments such as think tanks and research institutions. However, this view of a researcher is narrow and can exclude populations that have been historically underrepresented in these environments, such as people of color and economically disadvantaged people. More expansive definitions of who is a researcher take into account the work being done, rather than the context in which that work is carried out. Research is the systematic examination of a question to glean new knowledge or establish facts, and researchers are those carrying out this examination. Research can be done in a variety of contexts, from community advocacy groups to news rooms to more traditional contexts such as universities.

Conduct outreach upon publication

The benefits of public government data are realized only if public audiences are aware that the data is available. To build awareness and attract interest, agencies should incorporate into their engagement plans strategies for publicizing newly released data and increasing its external visibility.[76] The specific approach to release outreach will depend on the kind of data being released and possible use cases. For instance, if an agency is releasing a high-level data dashboard to inform decisions made by general audiences (e.g., information on schools or housing), the corresponding communications strategy could involve wide social media campaigns or public events.[77] Conversely, if an agency is releasing bulk data downloads that are most useful to researchers, the outreach should target academic institutions and other research centers, taking care to include researchers who may not fit a traditional academic profile, such as community-based nonprofits.[78]

COMMUNITY ENGAGEMENT IN PRACTICE IN LOUISVILLE, KY

The Sunlight Foundation partnered with the Louisville Metro Government's Open Data Program to attract more community interest in publicly available data.[79] Through a series of open house meetings, workshops, and hackathon events, the team raised public awareness and kickstarted new applications.[80] The city's data team monitored viewership and downloads of the public datasets to understand the impact of their engagement efforts and successfully attracted a diverse range of participants to engage with the data.

ENSURE RESPONSIBLE USE AND PRIVACY PROTECTION

Many of the potential downsides of publishing government data stem from the risks that it may compromise individual privacy or otherwise be used irresponsibly upon release. Agencies publishing data bear the responsibility for minimizing these risks as they prepare data for release, maintain and manage public data programs, and retire or phase out public datasets.

Protect individual privacy

Personal privacy considerations are paramount in any discussion of publishing individual-level data since the disclosure of personal data can cause serious harm if the data can be linked to individuals.[81] Consequently, government administrators should ensure that any data released includes robust protection of individual privacy, which, depending on the kind of data in question, may require technical or computational strategies.

Privacy concerns are amplified for sensitive data that could threaten individuals' safety or well-being if it is exposed or is prone to abuse by enabling discriminatory practices. Data on sexual orientation, immigration status, residential status, criminal or juvenile justice records, or experiences with domestic abuse are a few (nonexhaustive) examples. Therefore, public administrators should exercise serious caution if they do decide to publicly release sensitive kinds of data. Even data that appears to be unidentifiable may still be reidentifed if the deidentification was not done carefully or if an attacker has access to other data that provides additional information about people in the deidentified set. For example, AOL released a large dataset of user search queries that had been stripped of identifying information such as user names or IP addresses, but reporters were still able to identify the searches of Thelma Arnold, a 62-year-old from Georgia.[82] These and other examples of the harms of poor digital privacy protection have become a hallmark of contemporary online life.

Protecting privacy often requires striking the right balance between achieving the benefits of making information public and mitigating the harms.[83] In particular, three values can be in tension:

- **Privacy:** When releasing data publicly, people whose information is in the released data can be exposed to privacy violations, opening them up to financial or social harm. This risk is inherent in the release of any data that contains information about individuals, even if the dataset has been aggregated or ostensibly deidentified.

- **Accuracy:** Accuracy refers to how closely the released data hews to reality or to the original data. Reducing the accuracy of a dataset, such as by adding noise (for instance, adjusting the age of a given subject from 31 to 38, as described in more detail in Appendix I on page 26), can serve to protect the privacy of the data subjects, but it also can reduce the reliability of any insights gleaned from the data. For instance, if the goal is to determine rates of homeownership by age, data with added noise can produce results that are different from the original dataset, depending on the amount of noise that was added. Accuracy can also refer to the completeness of a dataset: Removing data from a set, such as by pulling out identifiers or limiting reporting for certain groups, reduces the accuracy of the data as it no longer represents the full context in which it was collected.

- **Granularity:** Granularity is how specific the data is to any one person in a dataset. Aggregating data reduces the granularity of the data. This lack of granularity helps to preserve the privacy of people in the dataset, but it can also reduce the utility or adaptability of the dataset, as aggregate data provides insights only about groups of people, rather than individuals, and may not allow for follow-up analyses.

Recent innovations in the field of privacy-enhancing technologies provide government agencies with tools to achieve the desired balance of trade-offs among privacy, accuracy, and granularity within a given dataset. This report presents a detailed overview of these technologies in Appendix I, but in short:

- **Redaction, shuffling and perturbation, and suppression** are techniques that limit the accuracy or completeness of a dataset to protect the privacy of the data subjects.

    - In **redaction,** specific sensitive or identifiable information such as Social Security numbers are removed before the dataset is released. Redaction is generally insufficient to protect privacy as data can often be easily reidentified through other quasi-identifiers (such as a combination of zip code, gender, and date of birth, which together can, by some evaluations, uniquely identify 87 percent of Americans) or through linkage to outside datasets.[84]

    - **Shuffling and perturbation** add noise to the data by switching specific information about certain data subjects or simply slightly altering random information in the dataset so that the dataset does not actually match to specific people. Both perturbation and shuffling reduce the accuracy of a dataset and are most effective at preserving privacy if the specific methodology is kept secret. This situation makes it challenging for researchers and others using the data to know how accurate or valuable their findings and conclusions from the data are.

- In **suppression,** data elements, statistics, or findings that are likely to be easily identified (such as data about a small number of minority students in a classroom) are removed from a dataset before it is released to prevent identification of those individuals. Unfortunately, this removal means that data about already marginalized populations is omitted, making identifying factors such as disproportionate impacts of policies harder if not impossible.

With these techniques, the goal is typically to balance accuracy and privacy. Striking that balance is challenging, and these datasets can be difficult to use for some purposes. They also can still allow reidentification, violating the privacy of the individuals whose data was released. Consequently, agencies should do a robust analysis of their data to evaluate the risk of reidentification prior to release. Additionally, since reidentification can be a result of combining datasets, agencies should occasionally review their released data to make sure it is still at low risk of reidentification.

- **Differential privacy** is a mathematical definition of privacy;[85] a system is differentially private "if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not."[86, 87] Differential privacy refers to a number of mathematical techniques for providing insights from a dataset without revealing any information about the data subjects. The mathematical underpinnings of differential privacy approaches can provide strong privacy guarantees, but the systems are often very complex to administer. Additionally, the approaches typically rely on setting a "privacy budget" that limits the amount of information that can be released, and setting a budget that makes for useful data release while still preserving privacy can be challenging.

- **Synthetic data** involves creating and releasing artificial datasets that mimic actual data but contain no data from real individuals. There are a number of approaches to creating synthetic data (including some based on differential privacy). Whatever approach is used, the creators of the synthetic data have to determine what qualities of the original dataset should be re-created, meaning that if the data is used in ways the creator did not expect, it can produce unreliable results. Agencies may provide a "validation server," which allows users of the dataset to verify their findings on the synthetic data against the actual dataset to help limit the concerns that a synthetic dataset will be inaccurate in unexpected ways.

Reducing the accuracy or granularity of a dataset can serve to protect the privacy of the people whose data is included, but doing so can limit the potential utility of the dataset. The intended use for a given data release will inform which qualities to prioritize to ensure that the data is serviceable for the intended use and that the most useful aspects of the data are preserved. Additionally, public administrators should understand the strengths, weaknesses, and resource requirements of these tools to determine which techniques to consider for a given data release plan. Even with the use of cutting-edge technology, particular categories of data, especially biometric and location data, are difficult or impossible to truly anonymize and therefore should not be released publicly.[88, 89] In these cases, an approach such as synthetic data that does not release actual data may be better suited. In any case, an important component of a public data release is determining which privacy protection technique is best suited to both the data itself and the intended use.

Avoid exacerbating bias and inequities

In addition to privacy concerns, data that reflects systemic biases or unjust power structures can be used in ways that inadvertently perpetuate these inequities if inherent bias is not identified and mitigated. This situation can occur if the data is used to train a predictive algorithm or otherwise drive agency decisions. In these scenarios, publishing biased data could do more harm than good if subsequent users of the data assume it to be fair and objective, thus masking social inequalities behind the appearance of impartial analytic processes.

USE OF BIASED DATA CAN PERPETUATE INEQUITIES

In 2019, education news outlet Chalkbeat conducted an evaluation and impact assessment of nonprofit organization GreatSchools' ranking system for public schools.[90] Chalkbeat's analysis found that GreatSchools' ranking system consistently ranked predominantly white, Asian, and affluent schools the highest and lower-income schools or those predominantly serving Black and brown students the lowest. The researchers noted that the reliance on historically biased public data sources, namely standardized test scores, accounted for the ranking system's disparities. The report found that the use of this data to inform school choice decisions could lead to further segregation.

Agencies should respond to these risks by assessing the biases within data to be published and clearly communicating them to external audiences at the point where the data is accessed. Publishing this information as metadata and contextual documentation is key to helping audiences understand appropriate and inappropriate secondary uses of the data. Contextual information that identifies assumptions made when processing the data, including how the agency has handled irregularities, missing or ambiguous elements, inconsistencies across multiple sources, and generalization assumptions, is also important in preventing invalid inferences or other irresponsible practices.[91] In some cases, biases within the data may make the data unfit for publishing altogether.

EVALUATE RESOURCE CONSTRAINTS

The process of making government data public requires a range of resources. Agencies should evaluate resource costs early in the planning process to assess whether, and how, data should be released.

Setting up processes and infrastructure for publishing government data will require time and effort[92, 93] to manage and oversee data governance systems as well as coordinate input solicitation and external engagement throughout the release process (discussed on page 17).*** Publishing data will also require specialized expertise to clean, structure, manipulate, and analyze the data to be released. The extent of this effort will depend on the characteristics of the data being published (complexity, cleanliness, need for manipulation, etc.)[94] and the format in which it will be released (as a bulk data download, through APIs, etc.).[95] Agencies should also distinguish between the effort required to initially set up the data publishing process and the effort that will be needed on an ongoing basis to maintain the public data and release new datasets.

In addition to labor resources, publishing government data requires appropriate infrastructure and computational resources to process the data, use privacy-enhancing technologies as needed, and host the data for public access. Once again, the resources required will depend on the kind of data being released and the agency's existing technical infrastructure — simple processing and analysis can be performed with standard spreadsheet-editing tools, but more complex operations may require specialized software such as statistical analysis tools or systems to host the data for external access. Likewise, some release formats require more data hosting and access infrastructure resources than others. Static data available for bulk download requires the least amount of infrastructure; dynamically generated files, interactive web portals, data visualizations, APIs, and database query systems all require more development effort due to their technical complexity.

If an agency determines that more robust technical resources or expertise are needed than it can provide internally, the agency may choose to work with an external vendor. Selecting a reputable and effective vendor will be critical to preserving the privacy of data subjects while still producing an effective dataset. Agencies should investigate any vendors before acquiring their products (more information about this process is discussed in Appendix I on page 29). If they do not feel equipped to evaluate vendors themselves, agencies may consider hiring a consultant or consulting with other agencies with more internal expertise.

*** Note: The resources discussed on this page presume that the data to be published already exists in some aggregate form within the agency prior to beginning the publishing process. If new reporting infrastructure is required to collect data in the first place — for instance, if a state agency wants to automate the aggregation and reporting of subsidiary local agencies — then additional resources will need to be designed and adopted. These considerations exceed the scope of this guidance.

CONCLUSION

Making government data publicly available can provide many benefits to external audiences and the agency releasing the data, but these benefits can be attained only when potential risks are identified and minimized. The report has discussed four key steps in this process: establishing data governance, engaging external communities, ensuring individual privacy, and evaluating resource constraints. The guidance presented in this report is not exhaustive but is intended to give a general landscape of salient issues when publishing government data with a privacy and equity lens in mind. Even if a situation does not allow for implementing all of the suggested practices at once, agency administrators can take initial steps to prepare to make more data publicly available.

APPENDICES

APPENDIX I: TECHNICAL APPROACHES TO PROTECTING PRIVACY

Introduction

A number of technological approaches can enable the release of public information and insights if privacy considerations prohibit a straightforward release of data. While these technologies can provide substantial benefits, they generally require some compromises as well, and not every technology will be suited to every use case. The policy needs of the release, such as privacy restrictions, level of granularity needed for the intended use, and intended audience, will dictate the technological approach the agency should take. Consequently, agencies will need to understand what the goal of releasing their data is, as different goals will be served by different technologies.

Trade-offs

Three particular values are often in tension when considering data for public release:

- **Privacy:** When releasing data publicly, people whose information is in the released data can be exposed to privacy violations, opening them up to financial or social harm. This risk is inherent in the release of any data that contains information about individuals, even if the dataset has been aggregated or otherwise deidentified.
- **Accuracy:** Accuracy refers to how closely the released data hews to reality or to the original data. Reducing the accuracy of a dataset, such as by adding noise (for instance, adjusting the age of a given subject from 31 to 38), can serve to protect the privacy of the data subjects, but it also can reduce the reliability of any insights gleaned from the data. For instance, if the goal is to determine rates of homeownership by age, data with added noise can produce inaccurate results. Accuracy can also refer to the completeness of a dataset: Removing data from a set, such as by pulling out identifiers or limiting reporting for certain groups, reduces the accuracy of the data as it no longer represents the full context in which it was collected.
- **Granularity:** Granularity is how specific the data is to any one person in a dataset. Aggregating data reduces the granularity of the data. This lack of granularity helps to preserve the privacy of people in the dataset, but it can also reduce the utility or adaptability of the dataset, as aggregate data provides insights only about groups of people, rather than individuals, and may not allow for follow-up analyses.

Reducing the accuracy or granularity of a dataset can protect the privacy of the people who have contributed to the dataset, but doing so limits its potential utility. The intended use for a given data release will inform which qualities to prioritize to ensure that the data is serviceable for the intended use.

This following section describes a number of different privacy-preserving technologies for making data available publicly and discusses the benefits, challenges, and risks of each approach.

Redaction, shuffling and perturbation, and suppression are techniques that limit the accuracy or completeness of a dataset to protect the privacy of the data subjects.

- **Redaction** is the removal of certain data from a dataset before it is released publicly, reducing the completeness and accuracy of the dataset. Typically data that is considered particularly sensitive or identifiable, such as names or Social Security numbers, is removed to protect the privacy of data subjects. This method is straightforward to implement but is typically insufficient to preserve the privacy of data subjects, as seemingly nonidentifying elements such as age, gender, or zip code can be used to identify people in the dataset, a process known as reidentification.[96, 97, 98]

- **Shuffling and perturbation** are techniques that alter components of a data element or data subject. In shuffling, components of one data subject are switched with components from another. For example, if the true dataset contains a 45-year-old male person who lives in the 87505 zip code and a 37-year-old nonbinary person who lives in the 02906 zip code, the publicly available may instead show a 45-year-old nonbinary person from 02906 and a 37-year-old male person from 87505. In perturbation, components of a data element are manipulated or perturbed to add noise to the set. For example, if the true dataset contains a 45-year-old male person, the released dataset might show a 57-year-old male person with all the same attributes. Both of these techniques reduce the accuracy of the dataset. These techniques are most effective at preserving privacy if the extent and type of the noise is not disclosed. This approach can make it difficult for researchers, journalists, advocates, or others using the data to know what sort of errors may be present in their findings and what the potential impacts of those errors may be.

- **Suppression** is the practice of removing or not reporting data from populations that are so small that they would be easily identified. For instance, consider a class of 30 students with only two Asian students. Those students would be easily identified by anyone with even passing knowledge of the class. A suppressed dataset would not include any information about those students, instead just reporting on the remaining 28 students. The released dataset may contain some indication that it is incomplete in some way. A significant limitation of suppression as a technique is that it may result in the omission of data from already marginalized groups. This situation can make it impossible for researchers to understand the impact of policies on these groups or examine disproportionate impacts.

With redaction, shuffling and perturbation, and suppression, the primary tension is between privacy and accuracy. Ideally, the accuracy is reduced enough to protect subjects' privacy but not enough to significantly affect insights drawn from the dataset. However, that balance can be difficult to strike. Further adding to this challenge, a dataset may seem to be sufficiently deidentified, but the addition of more outside data may add enough information to reidentify the dataset, so ensuring that a redacted or perturbed dataset will not be reidentified at some future point is difficult.

Differential privacy is a mathematical definition of privacy;[99] formally, a data processing algorithm is differentially private "if by looking at the output, one cannot tell whether any individual's data was included in the original dataset or not."[100] More generally, the term has come to refer to a number of mathematical and statistical techniques for providing insights from a dataset without revealing any information about the data subjects in the dataset. Because differential privacy refers to a collection of techniques, releasing differentially private data can take a few different forms in a public release context.

One approach is releasing insights from a dataset with a guarantee that even with all those insights, an attacker could not determine if any given person was in that dataset or learn anything about a specific person based on whether or not they were in the dataset. Alternatively, an agency could allow the public to ask queries of a dataset but ensure that the queries are answered in such a way that the querier cannot learn anything identifiable from the answers. This is done by setting a "privacy budget" that limits the amount and type of data that can be released. Each query draws from the privacy budget, and once the budget is spent, no more queries can be answered. For example, imagine that a member of the public is trying to learn about student outcomes in different schools. They may ask, "What was the average grade point average (GPA) for white students in 3rd grade at Jones Elementary School?" and receive the answer of 3.2. Then they ask how big the class is and get the answer that it is 25 students. They can then ask, "What was the overall average GPA for that class?" and get the answer 3.4. They may then ask, "How many non-white students are in the class?" The answer is that there is only one non-white student in the class. However, providing this information to the person asking the question would reveal the GPA of that student, so the question cannot be answered. Depending on the specific algorithm used, differential privacy approaches will sacrifice either accuracy or completeness to preserve the privacy of data subjects.

While the mathematical underpinnings of differential privacy provide assurance that the privacy of data subjects is protected, differential privacy approaches are typically very complex to administer, requiring significant data science expertise and resources. Additionally, setting a privacy budget that protects the privacy of subjects while still allowing for the release of enough data (and with enough accuracy) for meaningful transparency or research may not always be possible.[101]

Any approach that allows the public to query a dataset must also account for possible collusion between people querying the dataset. Going back to the previous example, imagine two different people are querying the dataset. The first person, Alice, asks the first three questions. Bob asks the fourth question. If all of their questions are answered, Bob and Alice can pool their information and learn more than they should have. Consequently, to maintain the privacy guarantees, the privacy budget must be reduced with every query into the dataset, regardless of who is asking (so everyone who queries the dataset is sharing one privacy budget).

Synthetic data is an artificial dataset that is meant to mimic the qualities of a true dataset without containing any data from a true subject, sacrificing accuracy in the name of privacy.[102, 103] So, if a true dataset has a 3rd-grade class, the synthetic dataset might have a 3rd-grade class with a similar size, demographic distribution, and average GPA but have no individual subjects with the same information as a true subject. Synthetic datasets can be constructed using a number of different techniques, including those based on differential privacy frameworks, offering the mathematical privacy guarantees conferred by these frameworks.

Regardless of the method of construction, creators of a synthetic dataset must determine which qualities should mirror the real dataset and which can vary. If the dataset is queried in unexpected ways or used for an unexpected purpose, the results may not be reliable. Additionally, it may be difficult or impossible to create a dataset that is sufficiently similar to the original to be useful but sufficiently different to provide privacy, particularly for small datasets.

Each of these approaches offers different benefits and drawbacks and will be suited to different purposes and contexts, meaning that any release using one of these techniques is most likely to be successful if the agency understands how the released data is likely to be used and by whom.

Vendor management

Each of the approaches listed in the previous section requires technical expertise and resources. Some agencies may choose to rely on vendors to prepare their data for release if they do not have the appropriate expertise and resources in house. In these cases, choosing an appropriate vendor that uses effective techniques will be critical to maintaining the privacy of data subjects. Agencies should investigate vendors' offerings before procuring their services.

- Vendors should be able to clearly articulate what approaches they take to maintain the privacy of the data subjects. Additionally, they should be able to describe why their approach is suited to the intended use.
    - What techniques does the vendor use to preserve privacy?
    - Why are those techniques suited to the given use?
- Vendors should be able to explain the benefits of their approach and also the challenges or drawbacks entailed by that approach. As preparing data for public release always involves balancing different priorities, no approach will be without shortcomings, so if a vendor is not able or is unwilling to discuss the trade-offs, agencies should consider that cause for concern.
    - What privacy guarantees are offered by the vendor's approach?
    - What are the drawbacks or risks of that approach?
    - In what cases will the approach "fail" (whether the failure is a breach of privacy or a dataset that is not useful)?
- For the creation of synthetic datasets, the vendor should be able to explain what techniques they use to create the dataset and what privacy guarantees that technique offers. Additionally, they should be able to explain what characteristics the synthetic dataset replicates from the original data and why those characteristics are valuable given the intended use.
    - What techniques does the vendor use to create the synthetic data?
    - What privacy guarantees are conferred by that approach?
    - What characteristics does the synthetic data set replicate?
    - Why were those characteristics chosen?
    - What characteristics will diverge from the original data?

Agencies may need internal expertise to evaluate the quality of vendor offerings and determine if the answers to these questions are sufficient for the agency to procure that vendor. If the agencies do not have internal expertise on hand, they may consider hiring a neutral consultant to help them evaluate the vendor product.

APPENDIX II: MACHINE-READABLE DATA RELEASE FORMATS

An important factor to consider when choosing a format for releasing data is the ease of machine-readability, one of the principles of open data.[104] Machine-readability refers to the ease with which an automated software program can ingest and process a data source without relying on manual customizations.[105] Machine-readability is essential for allowing scalable, stable data use in analytic applications (e.g., research projects), which would otherwise require prohibitively high human labor costs. Well-structured data in commonly used formats (such as CSV files) is the most machine-readable; proprietary, ad hoc, or poorly structured formats (such as PDF files) are the least machine-readable.

Machine-readability also includes the ease with which an automated program can access data hosted by a website. Datasets that are scattered throughout a website, or whose URLs are prone to changes, are less machine-readable because changes to the website structure can break automated data retrieval processes, requiring human intervention. Data that is held and accessed within a single, well-defined location through stable URLs is more machine-readable.

APPENDIX III: PUBLIC DATA LEGAL FRAMEWORKS

The following chart identifies examples of federal and state laws related to both public access to data held by governmental entities and privacy protections for public releases of that data. As described in more detail in the box on page 9, these examples highlight key trends in federal and state approaches to public data and the diversity of those approaches. This list of statutes, regulations, and guidance should be considered illustrative and is by no means exhaustive. Additional laws may apply to particular situations, and entities should always consult with counsel to determine their exact legal responsibilities.

Law/Authority	Type	Cite	Key Provisions
Federal Records Act of 1950, Pub. L. 81-754	Federal open data law	44 U.S.C. § 2108(a)	Requires federal agencies to convey records to the National Archives for preservation and public availability
Pub. L. 87-813 (1962)	Federal data privacy law	13 U.S.C. § 9	Prohibits the Census Bureau from using Census data "for any purpose other than the statistical purposes for which it is supplied"
Freedom of Information Act (FOIA), Pub. L. 89-487 (1966); Pub. L. 110-175 (2007); Pub. L. 114-185 (2016)	Federal open data law	5 U.S.C. § 552	Permits the public to request access to federal agency records or information, subject to nine exceptions
Privacy Act of 1974, Pub. L. 93-579	Federal data privacy law	5 U.S.C. § 552a	Prohibits disclosure of information about individuals except for "a purpose which is compatible with the purpose for which it was collected," subject to exceptions
Pub. L. 95-416 (1978)	Federal data privacy law	44 U.S.C. § 2108(b)	Codifies the 1952 letter of agreement between the Census Bureau and National Archives establishing the requirement that personal information collected by the Census remain private for 72 years[106]
Computer Matching and Privacy Protection Act of 1988, Pub. L. 100-503	Federal data privacy law	5 U.S.C. § 552a(o)	Requires written agreements between federal and nonfederal agencies to share federal data in matching programs

Law/Authority	Type	Cite	Key Provisions
Office of Management and Budget (OMB), Privacy Act of 1974: Final Guidance Interpreting the Provisions of Public Law 100-503, the Computer Matching and Privacy Protection Act of 1988 (1989)	Federal data privacy policy	54 Fed. Reg. 25818	Implements and clarifies requirements for federal agencies to share individual data in matching programs
E-Government Act of 2002, Pub. L. 107-347	Federal data privacy law; federal open data law	44 U.S.C. § 3601 et seq.	Requires privacy impact assessments for "developing or procuring information technology that collects, maintains, or disseminates information that is in identifiable form" and regulates the confidential use of statistical information, known as CIPSEA, updated in 2018 (see below)
OMB, Guidance for Implementing the Privacy Provisions of the E-Government Act of 2002 (2003)	Federal open data policy	M-03-22	Provides guidance on when agencies must conduct privacy impact assessments in collecting, maintaining, or disseminating individual information
Digital Accountability and Transparency (DATA) Act of 2014, Pub. L. 113-101	Federal open data law	31 U.S.C. § 6101 et seq.	Requires the U.S. Department of the Treasury to establish common standards for financial data provided by all government agencies and to expand the amount of data that agencies provide to a public website on federal spending
Open, Public, Electronic, and Necessary (OPEN) Government Data Act, Pub. L. 115-435 (2018)	Federal open data law	44 U.S.C. § 3506 et seq.	Requires federal agencies to make data publicly available online in a machine-readable format under an open license
Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2018, Pub. L. 115-435	Federal data privacy; federal open data law	44 U.S.C. § 3561 et seq.	Requires agencies to distinguish between statistical and nonstatistical data and limits the use of statistical data to statistical purposes

Law/Authority	Type	Cite	Key Provisions
OMB, Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance (2019)	Federal open data policy	M-19-23	Provides high-level guidance on the development of open data plans
California State Administrative Manual (2022)	State open data policy	Secs. 5160–5160.2	Requires state agencies to "[p]rioritize the use of open formats that are non-proprietary, publicly available, and that place no restrictions upon their use"
Florida House Bill No. 5301 (2019)	State open data law	Fla. Stat. § 282.0051	Requires Florida Digital Service to recommend open data technical standards
Massachusetts House Bill No. 3731 (2017)	State open data law	Mass. Gen. Laws ch. 7D § 4A	Permits the establishment of a state chief data officer, who shall "develop administrative directives to govern the use, storage, collection, and dissemination of data assets"
New York Executive Order No. 95 (2013)	State open data policy	n/a	Directs state agencies to maintain an open data website for all "publishable data"
Texas Information Resources Management Act	State open data law	Tex. Gov't Code § 2054.070	Directs the Department of Information Resources to establish a "central repository of publicly accessible electronic data"

ENDNOTES

1 A Catalog of Civic Data Use Cases, Data-Smart City Solutions (Apr. 28, 2022) [https://perma.cc/3XVC-Z7LN].

2 Behind the Scenes: HPD's Housing New York Open Data, New York City Department of Housing Preservation and Development (Sept. 18, 2019) [https://perma.cc/J77G-CX52].

3 Sam Gill, Indi Dutta-Gupta, & Brendan Roach, Boulder County, Colorado: Integrated Service Delivery, Data-Smart City Solutions (July 16, 2014) [https://perma.cc/FS56-UBK3].

4 Using Labor Market Data to Improve Student Success, Aspen Institute (2016) [https://perma.cc/FU6F-ZYC7].

5 Our History, International Open Data Charter (Accessed June 24, 2022) [https://perma.cc/LVV8-XEWL].

6 Open Data's Impact, The GovLab (Accessed June 24, 2022) [https://perma.cc/W27X-AFN3].

7 Jeni Tennison, Open Data Is a Public Good. It Should Not Be Confused with Data Sharing, The Guardian (May 12, 2014) [https://perma.cc/Z5DB-VKQX].

8 Beth Simone Noveck, Rights-Based and Tech-Driven: Open Data, Freedom of Information, and the Future of Government Transparency, Yale Human Rights & Development Law Journal (2017) [https://perma.cc/66NL-MHHX].

9 Katya Abazajian, Natalie Ward, Kell Crowley, & Insha Momin, Open Data for Economic Recovery, Beeck Center (June 2021) [https://perma.cc/5WW9-S2RA].

10 Brandon Vigliarolo, Consumer Privacy Study Finds Online Privacy Is of Growing Concern to Increasingly More People, TechRepublic (Oct. 1, 2021) [https://perma.cc/U728-M6SS].

11 Tim De Chant, Catholic Priest Quits After "Anonymized" Data Revealed Alleged Use of Grindr, Ars Technica (July 21, 2021) [https://perma.cc/D7CE-CYVV].

12 Rafi Goldberg, Lack of Trust in Internet Privacy and Security May Deter Economic and Other Online Activities, National Telecommunications and Information Administration (May 13, 2016) [https://perma.cc/A3HX-TMWG].

13 Domestic Violence and Privacy, Electronic Privacy Information Center (Accessed June 24, 2022) [https://perma.cc/6YBM-3DGJ].

14 Danielle Genet, Judge Whose Son Was Killed Speaks Out about New Bill to Protect Judges, ABC News (May 11, 2022) [https://perma.cc/5J3B-63NP].

15 The Annotated 8 Principles of Open Government Data, Open Government Working Group (Dec. 2007) [https://perma.cc/32K4-PYBJ].

16 Open Data Policy Guidelines, Open Data Policy Hub (Accessed June 24, 2022) [https://perma.cc/J3AN-WHLX].

17 Mark J. Headd, Open Data Guide (2016) [https://perma.cc/2542-6HGP].

18 See note 15.

19 Fostering Equitable and Complete Neighborhoods: Case Study, Sunlight Foundation — Roadmap to Informed Communities (Accessed June 24, 2022) [https://perma.cc/FGX8-48BQ].

20 Reforming Services for People Experinecing Homelessness, Sunlight Foundation — Roadmap to Informed Communities (Accessed June 24, 2022) [https://perma.cc/LDF2-2ULY].

21 Youth Risk Behavior Surveillance, Centers for Disease Control and Prevention (Aug. 21, 2020) [https://perma.cc/G5TX-LMSS].

22 Liz Schott, State General Assistance Programs Very Limited in Half the States and Nonexistent in Others, Despite Need, Center on Budget and Policy Priorities (July 2, 2020) [https://perma.cc/URK2-QD8L].

23 Victoria Kabak, Using Data to Combat Infant Mortality in Cincinnati, Data-Smart City Solutions (May 21, 2014) [https://perma.cc/4LER-VUNM].

24 Robin Jacob, Using Aggregate Administrative Data in Social Policy Research, Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services (Dec. 2016) [https://perma.cc/K9R7-67QU].

25 Mike Cassidy & Janet Currie, The Effects of Legal Representation on Tenant Outcomes in Housing Court: Evidence from New York City's Universal Access Program, Princeton University (Mar. 2022) [https://perma.cc/3XGZ-3UJF].

26 Universal Access to Legal Services Law, New York City Housing Court (Accessed June 24, 2022) [https://perma.cc/E7J8-LP34].

27 How a Program in NYC Made a 'Huge Difference' for Poor Tenants in Housing Court, Princeton University Department of Economics Communications (Apr. 25, 2022) [https://perma.cc/9RMJ-F9S8].

28 James M. Ferris & Nicholas P.O. Williams, Philanthropy and Government Working Together: The Role of Offices of Strategic Partnerships in Public Problem Solving, The Center on Philanthropy and Public Policy, University of Southern California (Nov. 2012) [https://perma.cc/MV72-GXJR].

29 Open Data Watch & Tom Orrell, Maximizing Access to Public Data: Striking the Balance Between "Open by Default" and Targeted Data Sharing, Open Data Watch & SDSN TReNDS (Mar. 4, 2019) [https://perma.cc/QZL3-PGM7].

30 Michael Christopher Jelenic, From Theory to Practice: Open Government Data, Accountability, and Service Delivery, World Bank (June 2019) [https://perma.cc/K97Q-33H9].

31 Lisa J. Huriash, This Free-Spending City Is Still at Risk for Fraud and Waste, Broward's Top Watchdog Warns, South Florida Sun Sentinel (Apr. 29, 2022) [https://perma.cc/2JXZ-DSEX].

32 Horus Alas, Report: Chronically Underfunded Public Health Programs Led to States' Uneven Pandemic Response, U.S. News & World Report (May 7, 2021).

33 Tara García Mathewson, New Data: Even within the Same District Some Wealthy Schools Get Millions More Than Poor Ones, The Hechinger Report (Oct. 31, 2020) [https://perma.cc/T2Z7-UASY].

34 Erika McCallister, Tim Grance, & Karen Scarfone, Guide to Protecting the Confidentiality of Personally Identifiable Information (PII), National Institute of Standards and Technology, U.S. Department of Commerce (Apr. 2010) [https://perma.cc/LN3Q-PYPN].

35 143 Million Compromised Social Security Numbers: Everything You Need to Know About the Equifax Hack, The Verge (Sept. 7, 2017) [https://perma.cc/47P4-LCQ6].

36 Alex Hern, New York Taxi Details Can Be Extracted from Anonymised Data, Researchers Say, The Guardian (June 27, 2014) [https://perma.cc/25CM-FXMF].

37 Andrew Keats, Juvenile Records Laws Must Be Reformed to Prevent Ongoing Racism, Juvenile Law Center (Aug. 7, 2020) [https://perma.cc/YS5R-7HZK].

38 Eric Durnell, Karynna Okabe-Miyamoto, Ryan T. Howell, & Martin Zizi, Online Privacy Breaches, Offline Consequences: Construction and Validation of the Concerns with the Protection of Informational Privacy Scale, International Journal of Human–Computer Interaction (Aug. 12, 2020) [https://perma.cc/Z35W-8TFA].

39 Jenny Wu, Handle with Care: Domestic Violence Safety Planning in the Age of Data Privacy Laws, Seattle Journal of Technology, Environmental & Innovation Law (May 7, 2021) [https://perma.cc/8AXJ-RXQ6].

40 Data Security Levels — Research Data Examples, Harvard University Information Security (Accessed June 27, 2022) [https://perma.cc/BZH9-8LKW].

41 Domestic Violence and Privacy, Electronic Privacy Information Center (Accessed June 27, 2022) [https://perma.cc/7U43-U4P6].

42 Andrew Perrin, Half of Americans Have Decided Not to Use a Product or Service Because of Privacy Concerns, Pew Research Center (Apr. 14, 2020) [https://perma.cc/G7PE-AP2W].

43 Rafi Goldberg, Lack of Trust in Internet Privacy and Security May Deter Economic and Other Online Activities, National Telecommunications and Information Administration (May 13, 2016) [https://perma.cc/RB7G-BFHE].

44 Luc Rocher, Julien M. Hendrickx, & Yves-Alexandre de Montjoye, Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models, Nature Communications (July 23, 2019) [https://perma.cc/XKC2-RU2M].

45 Natasha Lomas, Researchers Spotlight the Lie of 'Anonymous' Data, TechCrunch (July 24, 2019) [https://perma.cc/8MJ8-S4KB].

46 Luk Arbuckle, Aggregated Data Provides a False Sense of Security, International Association of Privacy Professionals (Apr. 27, 2020) [https://perma.cc/7ZK5-W4A4].

47 Michael Wines, The 2020 Census Suggests That People Live Underwater. There's a Reason, The New York Times (Apr. 21, 2022) [https://perma.cc/83K4-4L9X].

48 Rob Matheson, The Privacy Risks of Compiling Mobility Data, MIT News (Dec. 7, 2018) [https://perma.cc/Q4X3-L9N6].

49 Yves-Alexandre de Montjoye et al., Unique in the Crowd: The Privacy Bounds of Human Mobility, Scientific Reports (Mar. 25, 2013) [https://perma.cc/VPP8-XAXW].

50 K–12 Education: Discipline Disparities for Black Students, Boys, and Students with Disabilities, United States Government Accountability Office (Apr. 4, 2018) [https://perma.cc/666R-NX5T].

51 Larry Abramson, 'LA Times' Teacher Ratings Database Stirs Debate, NPR (Aug. 27, 2010) [https://perma.cc/4FNN-DDC2].

52 Valerie Strauss, LA Times Rates Teachers Again, Unfortunately, The Washington Post (May 9, 2011) [https://perma.cc/F8QJ-5EVV].

53 Gregory Ferenstein, The LA Times Trolls Innocent Teachers, TechCrunch (Nov. 10, 2013) [https://perma.cc/6MJK-3FM7].

54 Matt Barnum, The LA Times Teacher Ratings Helped the Academically Rich Get Richer, Chalkbeat (Aug. 5, 2018) [https://perma.cc/XT9Z-SA2A].

55 Innovation Defined: How Much Does Open Data Cost?, Bloomberg Cities Network (Accessed Aug. 16, 2022) [https://perma.cc/ZQ7W-RA7K].

56 Communicating the Value of Data Governance, Statewide Longitudinal Data Systems Grant Program (Dec. 2017) [https://perma.cc/26VL-DN8D].

57 Data Stewards, Data Stewards Network (Accessed June 28, 2022) [https://perma.cc/EQ5N-JXP3].

58 Stefaan G. Verhulst, Data Stewardship Re-Imagined — Capacities and Competencies, Data Stewards Network (Oct. 8, 2021) [https://perma.cc/BMB6-RWDN].

59 Elizabeth Laird & Hannah Quay-de la Vallee, Report — Data Ethics in Education and the Social Sector: What Does It Mean and Why Does It Matter?, Center for Democracy & Technology (Feb. 22, 2021) [https://perma.cc/6TDK-EGM8].

60 Single Agency Data Governance: Roles and Responsibilities, Statewide Longitudinal Data Systems Grant Program (Dec. 2019) [https://perma.cc/7UVQ-7JJQ].

61 Stefaan G. Verhulst & Andrew Young, Open Data Demand: Toward an Open Data Demand Assessment and Segmentation Methodology, The GovLab (Dec. 2018) [https://perma.cc/R7EL-MKRG].

62 Ben Green et al., Open Data Privacy, Berkman Klein Center for Internet & Society Research Publication (2017) [https://perma.cc/4BXG-HYJM].

63 Data Cleaning — DIME Wiki, The World Bank (Accessed June 27, 2022) [https://perma.cc/VZ93-GX8X].

64 Ben Green et al., Open Data Privacy, Berkman Klein Center for Internet & Society Research Publication (2017) [https://perma.cc/4BXG-HYJM].

65 Mark J. Headd, Open Data Guide — Adding Metadata (2016) [https://perma.cc/HX7Q-C7EA].

66 How to Open Up Data, Open Data Handbook (Accessed June 28, 2022) [https://perma.cc/A9E5-WW96].

67 Open Data Watch & Tom Orrell, Maximizing Access to Public Data: Striking the Balance between "Open by Default" and Targeted Data Sharing, Open Data Watch & SDSN TReNDS (Mar. 4, 2019) [https://perma.cc/E8YQ-GS9Q].

68 DC Data Policy — Mayor's Order 2017-115, Open Data DC, Government of the District of Columbia (Apr. 27, 2017, amended June 19, 2018) [https://web.archive.org/web/20220816185840/https://opendata.dc.gov/pages/data-policy].

69 Mohamed Sabt, Mohammed Achemlal, & Abdelmadjid Bouabdallah, Trusted Execution Environment: What It Is, and What It Is Not, IEEE International Conference on Trust, Security and Privacy in Computing and Communications (Aug. 2015) [https://perma.cc/TJ26-54LZ].

70 State Open Data Laws and Policies, National Conference of State Legislatures (Jan. 25, 2022) [https://perma.cc/9XTJ-BXU4].

71 Elizabeth Laird & Hugh Grant-Chapman, Report — Sharing Student Data across Public Sectors: Importance of Community Engagement to Support Responsible and Equitable Use, Center for Democracy & Technology (Dec. 2, 2021) [https://perma.cc/4SHD-WGK4].

72 Carrie Pomeroy, How Community Members in Ramsey County Stopped a Big-Data Plan from Flagging Students as At-Risk, Twin Cities Daily Planet (Feb. 20, 2019) [https://perma.cc/3BWU-TD56].

73 Elizabeth Laird & Hugh Grant-Chapman, Report — Sharing Student Data across Public Sectors: Importance of Community Engagement to Support Responsible and Equitable Use, Center for Democracy & Technology (Dec. 2, 2021) [https://perma.cc/4SHD-WGK4].

74 Ania Calderon, Publishing with Purpose: Introducing Our 2018 Strategy, Open Data Charter (Jan. 29, 2018) [https://perma.cc/FWE9-BSYL].

75 Nathan Zencey, Who's at the Popular Table? Our Analysis Found Which Open Data the Public Likes, Sunlight Foundation (Sept. 11, 2017) [https://perma.cc/UQZ5-SEW8].

76 Jennifer Angarita & City of Cambridge, Summer 2016 Open Data Collaboration Research Report, Cambridge Open Data (Aug. 2016, updated Oct. 14, 2016) [https://perma.cc/NV3P-39KQ].

77 Measuring the Impact of Community Engagement around Open Data, Sunlight Foundation — Roadmap to Informed Communities (Accessed June 28, 2022) [https://perma.cc/4D48-FX3P].

78 Open Data Engagement Guidance, Federal Enterprise Data Resources — resources.data.gov (Accessed June 28, 2022) [https://perma.cc/7P5K-SY95].

79 Louisville, KY: Using Partnerships to Enhance Open Data, Sunlight Foundation — Roadmap to Informed Communities (Accessed June 28, 2022) [https://perma.cc/3GTS-ZQJN].

80 Jonathan Jay, What Open Data Says About Post-Harvey Mosquito Threats (Sept. 1, 2017) [https://perma.cc/MM88-74R5].

81 See notes 10–13.

82 Paul Ohm, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Review (Aug. 13, 2009) [https://perma.cc/8JRK-NCFY].

83 Ben Green et al., Open Data Privacy, Berkman Klein Center for Internet & Society Research Publication (2017) [https://perma.cc/4BXG-HYJM].

84 Latanya Sweeney, Simple Demographics Often Identify People Uniquely, Carnegie Mellon University Data Privacy Working Paper (2000) [https://perma.cc/4CY9-WS98].

85 Cynthia Dwork, Differential Privacy, Automata, Languages and Programming — Lecture Notes in Computer Science (Eds. Bugliesi, Preneel, Sassone, Wegener) (2006) [https://perma.cc/NX2Z-3X9F].

86 Differential Privacy, Harvard University Privacy Tools Project (Accessed June 2022) [https://perma.cc/2BLC-8ZPX].

87 Cynthia Dwork & Aaron Roth, The Algorithmic Foundations of Differential Privacy, Foundations and Trends in Theoretical Computer Science (2014) [https://perma.cc/Q8K5-24BX].

88 Justin Banda, Inherently Identifiable: Is It Possible to Anonymize Health and Genetic Data?, The International Association of Privacy Professionals (Nov. 13, 2019) [https://perma.cc/5DE3-QGL9].

89 Yves-Alexandre de Montjoye et al., Unique in the Crowd: The Privacy Bounds of Human Mobility, Scientific Reports (Mar. 25, 2013) [https://perma.cc/VPP8-XAXW].

90 Matt Barnum & Gabrielle LaMarr LeMee, Looking for a Home? You've Seen Greatschools Ratings. Here's How They Nudge Families toward Schools with Fewer Black and Hispanic Students, Chalkbeat (Dec. 5, 2019) [https://perma.cc/ZDT5-5B42].
91 Data Documentation — DIME Wiki, The World Bank (Accessed June 27, 2022) [https://perma.cc/7V46-VJAB].
92 How to Plan and Budget an Open Data Initiative, United Nations Development Programme & Partnership for Open Data (Sept. 22, 2014) [https://perma.cc/5GNE-Q6EE].
93 Innovation Defined: How Much Does Open Data Cost?, Bloomberg Cities Network (Accessed Aug. 16, 2022) [https://perma.cc/ZQ7W-RA7K].
94 Omar Elgabry, The Ultimate Guide to Data Cleaning, Towards Data Science (Feb. 28, 2019) [https://perma.cc/KWC5-EMT7].
95 Joshua Tauberer, Bulk Data or an API?, Open Government Data: The Book (Aug. 2014) [https://perma.cc/DA3T-4LJ7].
96 Latanya Sweeney, Only You, Your Doctor, and Many Others May Know, Technology Science (Sept. 28, 2015) [https://perma.cc/3LNE-EUMA].
97 Arvind Narayanan & Vitaly Shmatikov, Robust De-Anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset), IEEE Symposium on Security and Privacy 2008 (May 2008) [https://perma.cc/2FBX-F843].
98 Salvador Ochoa, Jamie Rasmussen, Christine Robson, & Michael Salib, Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study (Aug. 2002) [https://perma.cc/8NMZ-6AZW].
99 Cynthia Dwork, Differential Privacy, Automata, Languages and Programming — Lecture Notes in Computer Science (Eds. Bugliesi, Preneel, Sassone, Wegener) (2006) [https://perma.cc/NX2Z-3X9F].
100 Differential Privacy, Harvard University Privacy Tools Project (Accessed June 2022) [https://perma.cc/2BLC-8ZPX].
101 Jaewoo Lee & Chris Clifton, How Much Is Enough? Choosing ∈ for Differential Privacy, Information Security Conference 2011 (Oct. 2011) [https://perma.cc/66XS-42XB].
102 Luke Rodriguez & Bill Howe, In Defense of Synthetic Data, FATES on the Web 2019 (May 2019) [https://perma.cc/5BZS-ZCFU].
103 Bill Howe et al., Synthetic Data for Social Good, Data For Good Exchange 2017 (Sept. 2017) [https://perma.cc/44VT-VNUR].
104 The Annotated 8 Principles of Open Government Data, Open Government Working Group (Dec. 2007) [https://perma.cc/ZD93-KDUT].
105 Dean Ritz, Understanding Machine-Readability in Modern Data Policy, Data Foundation (July 2020) [https://perma.cc/4EKX-XXV4].
106 The "72-Year Rule," U.S. Census Bureau (Jan. 2022) [https://perma.cc/SZ63-5V69].

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY