SUNLIGHT FOUNDATION

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

# GUIDELINES FOR OPEN DATA POLICIES

Version 3 | March 2014

*The Sunlight Foundation created this living set of open data guidelines to address: what data should be public, how to make data public, and how to implement policy. The provisions are not ranked in order of priority and do not address every question one should consider when preparing a policy, but are a guide to answer the question of what an open data policy can and should do in striving to create a government data ecosystem where open data is the default. Setting the default to open means that the government and parties acting on its behalf will make public information available proactively and that they'll put that information within reach of the public (online), without barriers for its reuse and consumption. Setting the default to open is about living up to the potential of our information, about looking at comprehensive information management and making determinations that fall in the public interest.*

## What data should be public

### 1. Proactively release government information online

Most government information disclosure laws and systems currently in place, including right-to-know, freedom of information and public records laws, are vehicles for reactive disclosure. Reactive disclosure means that a question has to be asked before an answer is given and that public information must be requested before it is disclosed. *Proactive* disclosure is the opposite. Proactive disclosure is the release of public information before an individual requests it. In the 21st century that means proactively putting new information online, where people are looking for it.

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

Open data laws provide an opportunity not just to update and improve access to information that is already open and/or public but also to specify that new data sets and records be collected and published. Similarly, policies should be specific about what “new” data can mean: in some instances, this provision can be used to require that that new data be created, collected and released for the first time.

## 2. Reference and build on existing public accountability and access policies

Open data policies should be informed by existing provisions ensuring access to government information. Strong open data policies build upon the principles embodied by existing laws and policies that defend and establish public access, often defining standards for information quality, disclosure and publishing. Examples of accountability policies include open meetings acts, open records acts, ethics standards, campaign finance regulation and lobbying disclosure laws, to name a few. Building on precedent from these policies and others can help strengthen new open data requirements and inform where policy updates or revisions are necessary that an open data policy can address.

Building on existing accountability and access policies can also help define the term "data" as it is used in an open data policy. Data, as it's defined in open data policies, can be seen as the next iteration of public records. Existing laws defining the scope of public records could be used as the touchstone for defining data to be released proactively online. Public records exemptions, however, should not be used to limit the scope of the definition of data. Open data policies that define data using the definition of public records should be able to adapt to changing definitions of public records. For that purpose, definition by reference would be stronger than definition by the copying of language, which would force updates in more than one place.

Another benefit of using existing access policies as a foundation for open data is that it can help ensure *legal* rights to information. Policies that already outline standards for access to information often create a legal right to that information, and this could be used to ensure the legal right to open data by extension.

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

## 3. Build on the values, goals and mission of the community and government

An open data policy can be pursued with the intent of realizing many different varieties of public good, including greater government transparency, honesty, accountability, efficiency, civic engagement and economic growth. An explicit statement of goals, values or intention can help clarify the outcomes that a government hopes an open data policy will help achieve. A statement of mission highlights both the general importance of open data and the specific importance of releasing information for that government's particular political context. In addition, the process of developing this statement may serve the democratic goal of increasing public participation by bringing together a wide range of stakeholders to explore the value of open data from their own perspectives.

## 4. Create a public, comprehensive list of all information holdings

For an open data policy to have a strong foundation, you first need to know what data you have—and so does the public. Governments should conduct an inventory of existing data early in the process of open data policy development in order for the government and other stakeholders to be aware of the full potential dimensions of data release. While defining total information holdings may be a complex undertaking, governments should conduct as comprehensive a review of existing data information as possible, with the inclusion of information holdings that may benefit from becoming structured data themselves.

The inventory should itself be made public. Publicly accounting for agency information helps ensure that information is managed to benefit the public interest, allows for common understanding of what data the government holds, and can create efficiencies among government departments. It empowers policymakers and administrators to determine whether information is being appropriately managed, and empowers the public oversight of those determinations. An individual or group should be charged with oversight of the inventory to ensure its ongoing maintenance and accuracy. To make the listing of data as useful as possible, such a list should also encompass data that may be viewed as sensitive or unlikely to be released (along with any other helpful context.) In addition to setting the stage

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

for meaningful public discussions around dataset release, an inventory process can provide a roadmap for creating ambitious timelines (see Provision 27) and identify whether new data may need to be collected.

## 5. Specify methods of determining the prioritization of data release

While open data policies ideally enable the online release of all public government information, the release of data may end up being a staggered process for practical reasons, such as insufficient funding or staffing. Governments should be clear about the range of potential methods that could be used in determining the priority-order of data release.

A variety of goals, actors and events can contribute to the determination of data set prioritization. Because of the traditional relevance of ethics concerns to open government policy, data which provides oversight of high-frequency areas for governmental ethics concerns serves the specific goal of achieving accountable government. Publishing data which is used in the process of creating public laws or rules, data related to specific legislative or executive policy initiatives, or data which is created incidental to a new policy or regulation serves the goals of civic engagement and transparency. The goal of satisfying public demand can be achieved both through a review of the existing volume of requests for government data and through a new solicitation of public comment. (Although direct public participation is important, it should not serve as the sole method of data set prioritization, because this mode of participation can inadvertently serve to reinforce the specific preferences of people who are already comfortable engaging the government.) Finally, given practical concerns, the cost of releasing individual data sets is likely to be used as an aspect of determining priority for release. While cost may be a factor in determining the priority of data release, it should be balanced against other prioritization methods in order to produce a truly useful collection of public data.

## 6. Stipulate that provisions apply to contractors or quasi-governmental agencies

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

Information that is gathered from the public using public funds should remain publicly-accessible, regardless of government decisions to delegate its management. The government often uses third party entities or contractors to handle, research or generate government information. Nonetheless, government decisions to employ outside contractors should not result in the public losing access to its own information. The scope of public information should be defined to include information managed by vendors of government services. Similarly, open data policy provisions should explicitly apply to quasi-governmental agencies and other similar actors, such as multi-state agencies, government-sponsored entities and publicly-funded universities. Where information is collected from or on behalf of the public using the government's legislative, regulatory or spending power, the public should retain presumptive access to that information.

To ensure that the public retains access to its data, provisions should added whenever possible to the existing procurement, contracting or planning processes requiring government contractors release government relevant information openly.

## 7. Appropriately safeguard sensitive information

A well-crafted open data policy is complementary to pre-existing legislation and directives about access to public information (see Provision 2 for more details), which means that it can integrate pre-existing public access law exemptions for information that is sensitive for privacy, security or other reasons. In addition, the nature of online access for bulk information can produce its own privacy, security and liability concerns. Individual-level data requires special scrutiny if it refers to private individuals who are not serving as government vendors. However, information that may provoke concern if released at the individual-level can often be released in aggregate and thereby provide some degree of public information and value. Any exemptions must be carefully crafted to exclude only the most necessary categories of information. Valid privacy and security concerns should be addressed through provisions that recognize the public interest in determining whether information will be disclosed or not. For example, rather than saying "information relating to X topic is exempt from disclosure," provisions should require

that "information relating to X topic is exempt from disclosure *if the potential for harm outweighs the public interest in disclosure*." Public interest here does not mean public attention, but instead refers to interests like democratic accountability, justice and effective oversight.

Any exemptions to data release should be crafted in a way that does not cut out access to information for researchers. Information that might be too sensitive for release to the public online can often be used by academic or nonprofit researchers who have agreed to protect sensitive information and not release it, except in aggregate form or in other ways that limit the potential for harm. This kind of release, with the research and insights it empowers, would benefit the interests of accountability, justice and oversight. Balance testing should still be used to ensure privileged information-sharing is not given priority over full public release when the public interest outweighs the potential for harm.

## How to make data public

### 8. Mandate data formats for maximal technical access.

For maximal access, data must be released in formats that lend themselves to easy and efficient reuse via technology. (See the Open Data Handbook, The Power of Information, 8 Open Government Data Principles, the 10 Open Government Data Principles and Open Government Data). This means releasing information in open formats (or "open standards"), in machine-readable formats, that are structured (or machine-processable) appropriately. Plainly, "open formats" refer to a rolling set of "open standards," often defined by standards organizations, that store information in a way that can be accessed by proprietary or non-proprietary software means. These formats exist across an array of data types; a common example cited is CSV in lieu of XLS for spreadsheets (the former being accessible via a wider variety of software mechanisms than the latter). "Machine-readability" simply refers to a format that a computer can understand. One step beyond machine-readable data is structured data (or machine-processable data), a format intended to ease machine searching and sorting processes. While formats such as HTML and PDF are easily opened for most computer users, these formats are difficult to convert the information to

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

new uses. Providing data in structured formats, such as JSON and XML, add significant ease to access and allow more advanced analysis, especially with large amounts of information.

## 9. Provide comprehensive and appropriate formats for varied uses

In addition to releasing information in formats that allow for the maximal amount of technical reuse, appropriate methods of distribution should be considered, to maximize the degree of access, use and quality of published information. For example, if a government report is most effectively distributed via a PDF format, but contains data elements that would be most digestible via a structured format, both the report and accompanying structured dataset should be released with relative referential metadata (see Provision 13). Similarly, options for bulk download should also recognize the strength of allowing for access to information in various formats. This degree of access and interaction allows citizens and government alike to get the most out of the data.

## 10. Remove restrictions for accessing information

To provide truly open access, there must be the right to reuse government information (explored in Provision 11) and no technical restrictions such as registration requirements, access fees and usage limitations, among others. Whether these technical restrictions have been specifically put in place (i.e., access fees) or are the accidental result of the choice of data format or software (i.e., usage limits or copyright restrictions), it is appropriate for an open data policy to address and remove these barriers to access. The aim should be to be to provide broad, non-discriminatory, free access to data so that any person can access information at any time without having to identify him/herself or provide any justification for doing so. Both open data policies and the Terms of Use (or Terms of Service) associated with government data should maximize the accessibility and use cases for data. While a disclaimer of warranties can be added to limit government liability, this mandate should pose no further restrictions, such as by limiting who or for what purposes the data be used.

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

## 11. Mandate data be explicitly license-free

If information is to be truly public, and maximally re-usable, there should be no license-related barrier to the re-use of public information. To be completely "open," public government information should be released completely into the worldwide public domain and clearly labeled as such. Opening data into the public domain removes barriers to information access, helps disseminate knowledge, aids in data preservation, promotes civic engagement and entrepreneurial activity and extends the longevity of the technological investments used to open information in the first place.

An open data policy must be explicit about this because copyright law varies from jurisdiction to jurisdiction. Moreover, while the U.S. Copyright Act explicitly does not include federal government works, it is silent on U.S. state and local government works. This, coupled with the additional complexities of copyright law (and ownership of various types of government data), mean special attention should be applied to all government data and the ease of its legal re-use. If the government data in question is not explicitly in the worldwide public domain, it should be given an explicit public domain dedication [such as the Creative Commons CC0 statement or a Open Data Commons Public Domain Dedication and License (PDDL)—both of which combine a waiver and a license].

## 12. Charge data-creating agencies with recommending an appropriate citation form

While failure to provide attribution for government data should never be actionable, users of government data should be encouraged to note the origin of data sources by accurately citing those sources. The practice of citing government data can be encouraged by having direct data managers develop model citations for their data sets. These model citations should both list key elements of the source's identity that would be required to effectively identify an individual data source and identify the unit of government which created or maintains the data. Where data users are actually transforming government data in some way, encouraging the proper citation of government

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

data will allow end users to distinguish between problems with government data quality and intermediary data quality by providing a clear route back to the original source of the data.

## 13. Require publishing metadata

Providing a common and fully described core metadata scheme (as well as other documentation) can be useful for the public and government alike. A strong metadata scheme takes its lead from common international meta attributes (such as DCAT), and allows data publishers to classify contextual fields or elements within their datasets. Commonly defined fields for such notations not only provide helpful context about the data's creation, quality and uses, but also help automate discovery mechanisms at the granular level, serving both government interoperability and the public discovery process.

## 14. Require publishing data creation processes

Providing quality data and insight into the operations of the government via government information requires an understanding of how the data was created. A summary of the processes that were used to create a specific data set provides valuable context that might not be discernable via metadata alone and should accompany the data set's release. Documentation of the workflow helps the public and government alike discern qualities about the dataset otherwise unavailable, such as (but not limited to): the sourcing, reliability, rarity and usability of the data. Additionally, documenting data creation processes can identify redundancy and areas for workflow and data creation improvements.

## 15. Mandate the use of unique identifiers

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

Unique identifiers are reference numbers used to identify unique individuals, entities or locations. The use of unique identifiers within and across data sets improves the quality and accuracy of data analysis. Without unique identifiers, some analyses can become difficult or impossible, since similar names may or may not refer to the same entities. Importantly, identifiers should be non-proprietary and public.

Several approaches could be taken to the development and dissemination of unique identifiers. For example, managers of individual data sets could be charged with developing the unique identifiers for the entities they most reference. Alternatively, a lead actor may oversee the development of a comprehensive identifier development schema. See also this list of extensive resources about the need for unique identifiers for corporate entities.

## 16. Require code sharing or publishing open source

In addition to data, the code used to create government websites, portals, tools and other online resources can provide benefits as valuable open data itself. Governments should employ open source solutions whenever possible to enable sharing and make the most out of these benefits.

## 17. Require digitization and distribution of archival materials

Open data policies can address not only information currently or soon to be available in an electronic format, but also undigitized archival material. Examples include everything from old budgets or meeting minutes to photos and maps.

Questions about what archival material should be digitized and what timelines are realistic for digitizing archival material can be informed by the same kind of prioritization process used for general data release (see Provision 5). Public participation and feedback from impacted government stakeholders will be key to making the digitization of archival material an effective process.

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

## 18. Create a central location devoted to data publication and policy

Data portals and similar websites can facilitate the distribution of open data by providing an easy-to-access, searchable hub for multiple data sets. At their best, these portals or hubs promote interaction with and reuse of open data and provide documentation for the use of information (see Provision 13). Portals can be generalized or specific (e.g., a spending or ethics portal), and can vary in terms of their sophistication. For specific portals, they should link to related portals when appropriate. Users looking at a portal for city campaign finance data, for example, could benefit from seeing a direct link to that city's portal for lobbying information. Portals and other related websites also provide governments with the opportunity to go into detail about issues and policies related to its commitment to openness and transparency. To facilitate their findability these websites should permit indexing and searching by third parties such as search engines.

There are several helpful features that should be included in general or specific portals. A list of what data is contained there is one necessary feature that makes it easy for users to quickly see what kinds of information are available on the data portal. If appropriate, this could be done through a link to a data inventory. Another beneficial feature to include in data portals is a view of analytics on data downloads. This will help users and government data providers understand what datasets are of the highest interest.

## 19. Publish bulk data

Bulk access provides a simple but effective means of publishing data sets in full by enabling the public to download all of the information stored in a database at once. This is a step beyond simply making select data sets or search results available for download or export and is critical for supporting the maximal reuse and analysis of data. Whether offered as a feature of a data portal—or even as a simple “click to download” button on a government

agency webpage describing or displaying information—bulk access to information is often one of the simplest and most direct steps a government entity can take to share public information.

## 20. Create public APIs for accessing information

Although bulk data (see Provision 19) provides the most basic access to searching and retrieving government data, government bodies can also develop APIs, or Application Programming Interfaces, that allow third parties to automatically search, retrieve or submit information directly from databases online (see Open Government Data). Navigating requirements for bulk data and APIs should be done in consultation with people with technical expertise as well as with likely users of the information.

## 21. Optimize methods of data collection

To optimize data quality and timeliness, disclosure regulations should take advantage of online data-collection methods. Electronic filing, also known as "e-filing," is one method of optimizing the quality and timeliness of data collection. To avoid the inefficiencies created by paper-based filing systems, governments should require online, electronic filing so long as filers can be reasonably expected to have access to the necessary technology. Electronic filing requirements save money, make real-time disclosure possible and allow structured data to be created at the same moment information is being filed, whereas paper filings only make reuse and analysis more difficult. Electronic filing provisions should include detailed language about what constitutes a “complete” filing and what to do if the online e-filing service is down.

## 22. Mandate ongoing data publication and updates

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

The ideal of online data is “real time” access: data should be made available as close as possible to the time that it is collected. It is not enough to mandate the one-time release of a data set because it becomes incomplete as soon as additional data is created but not published. In order to ensure that the information published is as accurate and useful as possible, specific requirements should be put in place to make sure government data is released as close as possible to the time that it is gathered and collected.

While sometimes challenging, this kind of rapid publishing becomes less of a burden when combined with others measures for digitizing data collection and publishing, such as electronic filing (see Provision 21), central data locations (see Provision 18) and APIs (see Provision 20).

### 23. Create permanent, lasting access to data

Once released, digitized government data should remain permanently available, “findable” at a stable online location or through archives in perpetuity. Although portals and websites can be vehicles for accessing this data over the long term (see Provision 18), it is critical that the data’s permanent release and accessibility is defined so as to apply to the data itself, not just the means of access.

Provisions relating to permanence can also be expanded to relate to updates, changes or other alterations to the data. For best use by the public, these changes should be documented to include appropriate version-tracking and archiving over time. These provisions should build on the strengths of existing records management laws and procedures (see Provision 2).

## How to implement policy

### 24. Create or appoint oversight authority

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

Some questions may defy easy treatment in the process of creating an open data policy, so it's appropriate to define a single authority empowered to resolve conflicts and ensure compliance with new open data measures. Some policies direct a pre-existing officer (e.g., a chief data or information officer, or an open data ombudsman) or a specific department to oversee execution and compliance, although new positions and authorities can also be created. It's important to emphasize that creating oversight does not necessarily require hiring new staff. Responsibility can be distributed among departmental coordinators who meet regularly, for example, to reduce the burden of oversight. This can also help with cross-departmental coordination and buy-in to the open data efforts.

Specifying an authority, review board or similar body is an important step to making sure that an open data policy can be executed and provides a resource to address unforeseen hurdles in implementation. Oversight bodies should conduct their work independently and publicly, and can be bolstered by creating new regulations or guidance for implementation (see Provision 25).

## 25. Create guidance or other binding regulations for implementation

Open data policies should be practically aspirational, meaning that they should define a vision for why the policy is being implemented, but also be able to provide actionable steps for the government and oversight authorities to follow to see the policy through to implementation. Creating regulations or guidance can ensure a strong, reliable policy and usually mean the difference between policy passed for show versus policy passed for substance. Regulations help make the work of oversight and implementation authorities possible. Open data policies can also direct guidance to be created from a basic framework described in the policy. So, rather than spelling out the entirety of data standards in the original policy document, governments can include direction in their policies that guidance be created to help agencies comply with online public access to non-proprietary, machine-readable data published in open formats.

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

## 26. Incorporate public perspectives into policy implementation

Just as public preferences should be incorporated into the development of an open data policy and in the prioritization of data sets for initial release, the public should be involved in the ongoing assessment and review of the policy's implementation. Governments should create meaningful opportunities for public feedback about data quality, quantity, selection, and format, as well as the user-friendliness of the point of access. In addition, this feedback should be formally considered and addressed when the policy undergoes review (see Provision 31).

## 27. Set appropriately ambitious timelines for implementation

Setting clear deadlines can demonstrate the strength of a commitment and will help translate commitments into results. Deadlines can also help identify failures clearly, opening the door to public oversight. Relevant actors should be given enough time to prepare for the changes brought on by the new open data policy, but not so much time that the policy becomes inoperable. The timeline should be firm, provide motivation for action and have actionable goals and benchmarks that can be used as a metric for compliance. These goals or checkpoints can include qualitative and quantitative measurements.

## 28. Create processes to ensure data quality

Data quality will not be ensured through data release alone: efforts need to be made to keep the data up-to-date, accurate and accessible. Data release should be approached as an iterative and ongoing process. As soon as sensitive information and security concerns are met, data should be released and regularly updated as it improves and grows. Data with serious accuracy and quality concerns should be adequately documented to avoid creating confusion or misinformation. Similarly, public data reporting streams that are separate from what is used within government should be avoided whenever possible, as redundant or parallel data streams can create opportunities for data quality to suffer. Each update should include clear and complete metadata (including a conspicuous contact person), group datasets where appropriate, and address concerns noted via a prominent feedback mechanism.

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

## 29. Ensure sufficient funding for implementation

Like any other initiative, implementing an open data policy should be done with an eye on long-term sustainability. One way to do this is to consider funding sources for the implementation of the policy as well as its future maintenance. Sufficient funding can mean the difference between successful and unsuccessful policies. Funding should be considered for—but not limited to—the potential of the following: new staff (administrative, technical and legal), new software (to house, extract and input data), training and server maintenance. While each jurisdiction's ability to fund will vary, significant consideration should go into identifying the resources reserved to assist and support an open data ecosystem.

## 30. Create or explore potential partnership

Partnerships can be useful in a variety of important efforts related to data release, such as: increasing the availability of open data, identifying constituent priorities for data release (see Provision 5), and connecting government information to that held by nonprofits, think tanks, academic institutions and nearby governments. Such partnerships can aid civic participation, help identify the gaps in service delivery, among other benefits. Public-private partnerships can be via contract, informal cooperation, or an exchange for rights or privileges. In addition to using commonly used formats, reaching out to nearby governments to explore ways to share data, experience, and workloads can assist in achieving open data outcomes.

## 31. Mandate future review for potential changes to this policy

Just as publishing open data is an ongoing process that requires attention to its quality and upkeep (see Provision 28), so too does the policy that establishes it. In order to keep up with current best practices and feedback from

1818 N Street N.W. · Suite 300 · Washington, D.C. 20036 ¦ 202.742.1520 ¦ F: 202.742.1524 ¦ SunlightFoundation.com

existing policy oversight, open data policies should mandate future review of the policy itself as well as of any guidance created by the policy or other implementation processes.

Open data policies should acknowledge that the context in which they operate is rapidly changing over time and will likely need sustained attention to remain relevant. There is a wide array of topics a review could, and should, cover. One key focus of review is understanding the audience for open data. Attention should be given to capturing details such as who is using government data, which data is being used, what the data is being used for and more.