Preserving State Government Digital Information
Minnesota Historical Society

Best Practice Principles for Opening Up Government Information

Summary

Users want access to government information and content. Access to data empowers people to learn, share, reuse, mashup, and create individualized or specialized applications that give the information new use value.

The Minnesota led NDIIPP project has been researching the best ways to facilitate access to government and legislative data. The following best practices principles help facilitate open government, transparency, and accessibility.

DISCLAIMER:
This white paper is a topical overview and is in no way intended to offer legal advice. Consult an attorney for assistance with specific concerns or for advice.

Any comments, corrections, or recommendations may be sent to the project team, care of:

Carol Kussmann
Collections Assistant, State Archives
Minnesota Historical Society
carol.kussmann@mnhs.org / 651.259.3262

Introduction

The Sunlight Foundation[1] and others, including the general public, have long thought that governments should be more transparent. The Federal Government tackled this issue when it created the Open Government Initiative[2] in order to help "establish a system of transparency, public participation, and collaboration"[3]. This initiative includes providing access to government online.

Access to data empowers people to reuse, mashup, share, and learn from available information and created applications. Success of websites such as data.gov and opencongress.org show the value in making data and resources available for use.

- Data.gov[4]: Provides datasets in machine readable formats from many data sources for research and analysis. Developers have created mashups and applications for public analysis and use with this data.

[1] Sunlight Foundation. *Home Page*. February 22, 2011. http://sunlightfoundation.com/
[2] Open Government Initiative. *Home Page*. January 6, 2011. http://www.whitehouse.gov/open
[3] President Obama. *Open Government Initiative: Home Page*. January 21, 2009. http://www.whitehouse.gov/open
[4] Data.Gov. *Home Page.* http://www.data.gov/

- OpenCongress[5]: A non-profit, non-partisan public resource for information on U.S. Congress bills, legislators, votes, issues, committees, and money in Washington, D.C. that uses publicly available information.

Many government agencies have a goal to become more transparent, to provide data to the public, and become more accountable but don't know where to start. Expanding on a list of principles produced by open government advocates including the Sunlight Foundation, the following principles provide a lens to evaluate the extent to which government data is open and accessible to the public.

Best Practice Principles

Completeness

Datasets released by the government should be as complete as possible, reflecting the entirety of what is recorded about a particular subject. All raw information from a dataset should be released to the public, except to the extent necessary to comply with federal law regarding the release of personally identifiable information. Metadata that defines and explains the raw data should be included as well, along with formulas and explanations for how derived data was calculated. Doing so will permit users to understand the scope of information available and examine each data item at the greatest possible level of detail.

Primacy

Datasets released by the government should be primary source data. This includes the original information collected by the government, details on how the data was collected and the original source documents recording the collection of the data. Public dissemination will allow users to verify that information was collected properly and recorded accurately.

Timeliness

Datasets released by the government should be available to the public in a timely fashion. Whenever feasible, information collected by the government should be released as quickly as it is gathered and collected. Priority should be given to data whose utility is time sensitive. Real-time information updates would maximize the utility the public can obtain from this information. Citizens would be better able to influence the public process and journalists would be able to cover stories in a more timely fashion.

Ease of Physical and Electronic Access

Datasets released by the government should be as accessible as possible, with accessibility defined as the ease with which information can be obtained, whether through physical or electronic means. For content available online, the content itself should be easy to find and download.

[5] Open Congress. *Home Page*. February 21, 2011. http://www.opencongress.org/

Barriers to physical access include requirements to visit a particular office in person or requirements to comply with particular procedures (such as completing forms or submitting FOIA requests).

Barriers to automated electronic access include making data accessible only via submitted forms or systems that require browser-oriented technologies (e.g., Flash, Javascript, cookies or Java applets). This includes having to enter in search terms to find data.

By contrast, providing an interface for users to download all of the information stored in a database at once (known as "bulk" access) and the means to make specific calls for data through an Application Programming Interface (API) make data much more readily accessible.

Machine readability

Machines can handle certain kinds of inputs much better than others; available datasets should be able to be read by machines easily and without introducing errors. For example, information shared in the widely-used PDF format, is very difficult for machines to parse. Handwritten notes on paper are very difficult for machines to process and scanning this text with Optical Character Recognition (OCR) results in many matching and formatting errors.

To reduce introduced errors, information should be stored in widely-used file formats that easily lend themselves to machine processing. If other factors necessitate the use of difficult-to-parse formats, data should also be available in machine-friendly formats and these files should be accompanied by documentation related to the format and how to use it in relation to the data.

Suggested machine readable formats, in order of preference on ease of use are:

- Structured
  - JSON, CSV, XML
  - Other formats: SQL Dump, DBF, Excel, etc.
- Semi-Structured: HTML/Plain Text
- Worst Case: PDF only

Authentication

Government data and records have long been shared with and made accessible through third parties, such as archives, libraries, legal publishers and news organizations. The advent of new and digital technologies has increased the number of such parties that are interested in and capable of presenting such content; new technologies and new organizations mean new practices, new policies and new questions about the authenticity and trustworthiness of the information they represent.

One established evidentiary principle to apply is the “chain of custody,” which would imply the ability to provide or reconstruct the history of exchanges and transformations, all inevitable and necessary for digital content, from the creation of a record to its

delivery and use. How much documentation would be required will probably vary from situation to situation; the needs of a casual researcher will be different from those of a prosecutor.

To authenticate the data it uses in the Open States project, Sunlight records the URL of the data source and creates time stamps recording when and where the content was acquired. As many state legislatures do not use permanent URLS for their web resources, some of these links may break. To address that, Sunlight is considering additional means of creating and sustaining audit trails.

Use of Commonly Owned Standards

Commonly owned (or "open") standards are standards that are not owned by any company or individual. Using open standards can increase the number of users by removing costs associated with using proprietary software.

For example, if only one company manufactures the program that can read a file where data is stored, access to that information is dependent upon use of the company's processing program. Sometimes that program is unavailable to the public at any cost, or is available, but for a fee. A common example is Microsoft Excel, a fairly commonly-used spreadsheet program, which requires users to purchase a license.

Freely available or open software and formats often exist by which stored data can be accessed without the need for a software license. Removing this cost makes the data available to a wider pool of potential users.

JSON, XML, and HTML are examples of open formats that are readable by any web browser and other tools that come without any fear of patent or vendor lock-in. The Open Document Foundation[6] supports document formats similar to Microsoft products without the licensing issue.

Permanence

The capability of finding information over time is referred to as permanence. Information released by the government online should be available online in perpetuity. Often times, information is updated, changed or removed without any indication that an alteration has been made. Or, it is made available as a stream of current data, but not archived anywhere. For best use by the public, information made available online should remain online, with appropriate version-tracking and archiving over time.

This is especially true with legislative data that is session sensitive. If your site is redesigned each session, there is no reason to lose (or remove) data from previous sessions.

[6] The Document Foundation. *Home Page*. http://www.documentfoundation.org/

Another key aspect to permanence is that URLs do not change without providing a redirect to the new address. This reduces broken like error messages and lost data. Tim Berners-Lee explains this in detail in "Cool URIs Don't Change."[7]

Licensing

Maximal openness includes clearly labeling public information as a work of the government and available without restrictions on use as part of the public domain. The imposition of "Terms of Service," attribution requirements, restrictions on dissemination and so on act as barriers to public use of data.

Non-discrimination

"Non-discrimination" refers to who can access data and how they must do so. Barriers to use of data can include registration or membership requirements. Another barrier is the uses of "walled garden," which is when only some applications are allowed access to data. At its broadest, non-discriminatory access to data means that any person can access the data at any time without having to identify him/herself or provide any justification for doing so.

Usage Costs

One of the greatest barriers to access to ostensibly publicly-available information is the cost imposed on the public for access--even when the cost is minimal. Governments use a number of bases for charging the public for access to their own documents: the costs of creating the information; a cost-recovery basis (cost to produce the information divided by the expected number of purchasers); the cost to retrieve information; a per page or per inquiry cost; processing cost; the cost of duplication, etc.

Most government information is collected for governmental purposes, and the existence of user fees has little to no effect on whether the government gathers the data in the first place. Imposing fees for access skews the pool of who is willing (or able) to access information. It also may preclude transformative uses of the data that in turn generates business growth and tax revenues.

Specifics for Legislative Data

As requested by the public, Sunlight Labs is now working on a state-level application of OpenCongress called OpenGovernment[8]. Co-developed by Sunlight and the Participatory Politics Foundation with countless volunteers from many different states, this system is free to use and built on an open-source platform. Currently five states are live on the website, with more in development. While building the application and collecting data from all fifty states, developers identified key areas that many states failed to meet the above principles. These include:

Completeness of Data

[7] Berners-Lee, Tim. *Cool URI's Don't Change*. W3C Style. http://www.w3.org/Provider/Style/URI.html
[8] Open Government. *Home Page*. http://opengovernment.org/home

By far, the most important thing is that all relevant data is made available. Essentially every state makes the basics available: current legislator information, bill statuses, votes, committee memberships, etc. However, Sunlight has identified several commonly omitted items that are important for creating a full understanding. These include: historical legislator data, individual legislator votes, committee votes, and legislator IDs that are unique, permanent, and persistent between sessions.

Data Formats

Once all data is exposed, the next goal should be to provide it all in at least one well-known, machine-readable format. Preferably, data should be made available in bulk as that is the most common use case and will save server time/bandwidth in the long run. An API allowing access to individual bills, etc. is nice, but not as important or accessible as bulk access. Suggested formats, in order of preference on ease of use are:

- Structured
  - JSON, CSV, XML
  - Other formats: SQL Dump, DBF, Excel, etc.
- Semi-Structured: HTML/Plain Text
- Worst Case: PDF only

See the XML provided by Thomas.gov[9] for an example of a structured format used in OpenCongress.gov.

Additional Suggestions

- If possible, datasets should include categorization on items such as actions, votes, bills, sponsorships.
  - For example adding a tag to a vote indicating it is of type 'passage' is useful. Many states use legislative jargon that may mean something to close observers of the legislature, but confuse others. You would not want to replace the term 'third reading' with "passage", but augmenting the data in a way to indicate that third reading votes can be considered a passage vote is helpful to the novice interested in the data.

- Provide an easy way on your website to find out what has changed since a given date.
  - For the average user, finding out what has recently changed is a common desire. It is important for legislative websites to be able to answer the question "what has the legislature done since I was last on this site?"
  - For developers, being able to query a site for updates since a certain date rather than crawl the entire site each time reduces the load on the server and avoids redundant requests.

- Splitting up legislators' names into First, Middle, and Last assists people in identifying a legislator, as sometimes it is hard to tell the first, middle, and last name apart. Splitting these into separate fields also organizes data.

[9] The Library of Congress. *THOMAS: Home Page*. http://thomas.gov/

- Provide access to legislative journals. These journals provide context around votes, and other actions. Being able to link votes to the relevant sections of a journal can be useful to a citizen looking to learn more about the workings of the legislature.

Acknowledgement

Thank you to Sunlight Foundation and Sunlight Labs for allowing the reuse and expansion on the "Ten Principles for Opening Up Government Information"[10] and "State Best Practices"[11].

[10] Sunlight Foundation. *Ten Principles for Opening Up Government Information.* August 11, 2010. http://sunlightfoundation.com/policy/documents/ten-open-data-principles/

[11] Open States. *StateBestPractices.* August 12, 2010. http://code.google.com/p/openstates/wiki/StateBestPractices