

Digital Curation Issues Involving Open Government Data

Joshua Chalifour

www.chalifour.org

www.chroknowlogy.ca

www.knowfore.com

6 April 2015

Open data is a well-defined concept but in the public sector, there is some difficult work ahead for its digital curation.

Although the support and production of open data from governments around the world varies (with many not yet supporting it at all) there are clear movements to encourage and grow open government data initiatives. Within the realm of governments that *do* support and produce datasets open to the public, benefits that would otherwise accompany the availability of this open data are sometimes hampered due to incomplete adoption of best practices.

I'd like review some of the tenets of open government data, then I'll discuss some of the digital curation issues that are important to deal with for the success of open government data initiatives.

Tenets of Open Government Data, in Brief

The Open Definition (OD) (<http://opendefinition.org>) states that “Open data and content can be freely used, modified, and shared by anyone for any purpose.” The Open Definition, a project of the Open Knowledge Foundation (OKF), is frequently referenced in literature and by open data and open government projects. It is a foundational definition for understanding the principles of open data and open government initiatives. The Open Data Handbook (<http://opendatahandbook.org>) builds on the OD, arguing that interoperability (the mixed usage of different datasets) is the key feature of what makes open data important. Specifically, the mixing of different datasets allows people to develop greater insight from open data. To this end, the Open Data Handbook prioritizes the importance of availability and accessibility, reuse and redistribution, and universal participation.

Open government data shares the foundational OD of open data and then builds on it to include characteristics that are specific to government institutions, public policy, and the rights of the populace. The OKF has a working group devoted to open government data, which (in its own description of open government data) requires this data to conform with the Open Definition in addition to describing it as “Data produced or commissioned by government or government controlled entities.” (<http://opengovernmentdata.org>).

The OKF's Open Government Data project argues that government data should be open for the following three reasons. First, to support a democratic society citizens need to know and be able to share what the government is doing, thus open government data supports transparency. Second, open government data can be used to stimulate economic and social innovation. Third, open government data enables people to contribute (not just be informed) to society—an argument for participatory governance.

Many open government data ideas evolved from a 2007 working group's statement of eight principles for determining whether government data may be considered open or not (<http://opengovdata.org>). These eight principles address completeness, primacy,

timeliness, accessibility, encoding for machine processing, availability (includes anonymity and lack of discrimination), formats (specifically, they cannot be solely proprietary or allow exclusive control by a single entity), and finally unrestricted by regulatory licensing (though certain exceptions can be allowed to address issues such as privacy).

In the USA, the Sunlight Foundation built upon these principles to create open data policy guidelines, which help government organizations determine how to make data public, identify which data are appropriate to make public, and how to put policy in place for this process of opening up data. The Sunlight Foundation's position is strongly in favour of managing public information so that it defaults to open—its guide aims to support that process.

The Open Data Institute (ODI) published its Open Data Maturity Model, which recommends assessing “...data management processes, knowledge and skills, customer support and engagement, investment and financial performance and strategic oversight...”¹ so that public organizations can determine how well they adhere to best practices for publishing and using open data.

Although Canadian legislation related to open government dates back to at least the 70s, our Open Government Initiative started in 2011 by focusing on open information, data, and dialogue. The Canadian government maintains an open data portal that as of 2012 had over 272,000 datasets from 20 departments². Among other technological issues, the Canadian action plan highlights the priority of launching a virtual library, serving as the repository for government-published documents and data. Other countries have been working on similar initiatives—in fact the Open Government Partnership (OGP) (<http://www.opengovpartnership.org>) helps governments establish their open data policies and initiatives. It provides guidance, which allow governments to structure their initiatives toward commonly accepted open government data goals and benchmarks the status of different governments' progress.

The European Commission's studies found that as of 2014, citizens had a difficult time finding and reusing open government data. Part of the EC's efforts to alleviate this problem were to recommend standard licences that did not restrict re-use and place as few restrictions as possible on datasets. It emphasized that usage requirements ought to be explicit.

The G8 Open Data Charter commits member countries to developing their open data initiatives in accord with certain principles. It recognizes that people want to access electronic data and services whenever and however they like³. It recognizes social, governance, and economic benefits that accompany open data. The plan commits members to releasing data high in both quality and quantity. It further stipulates a number of technical requirements, including to have a thorough plan for metadata maintenance, use open formats, reveal data standards, use open licences, ensure machine readability, and use application programming interfaces (APIs) where possible. These are all in-line with

preexisting best practices and principles of open data and the open definition itself.

The Open Data Working Group of the OGP drafted an inventory of technical standards for open government data⁴, which looked at catalogue structures for open data, metadata used (including controlled vocabularies, federation, and elements, file formats, and licences. A study of the systems used to enable open data found many varieties, which need to interact within an ecosystem. The ecosystem is a group of interdependent technical and social systems, which “...stimulates the participation of citizens in governmental processes of decision making and policy making.”⁵ The ecosystem includes tools for data storage and curation, open data catalogues, open government portals, tools for making requests, discussion areas, and analytics each of those correspond with processes by the data provider (government organization) and activities from the users.

Digital Curation Concerns

Several digital curation concerns exist with open government data. Among these are maintenance and standards, both of which impact accessibility. Referring to the role of records managers with open government data, Julie McLeod⁶ felt they should work with systems designers and data creators to ensure that what is developed (with respect to systems and policies) and provided treats metadata properly and supports interoperability as well as legitimate accessibility. These provisions relate directly to ongoing maintenance but they also ensure that data truly comply with the open definition.

Maintenance

The importance being put on organizing and releasing government data to the public, presents a sizeable challenge for ongoing digital curation. There is the effort to ensure that appropriate systems are developed and maintained to support data preservation but also some strategic concern for intra- and inter-governmental agency coordination.

Systems must support recognized digital preservation issues such as trust, authenticity, and reliability, as well as contending with some special use contexts. Namely, there is a motivation to make government data open because it will increase innovation, public participation in governance, and other beneficial objectives.

As noted by the Open Data Handbook, interoperability is key. With that in mind, those involved in digital curation have to consider the ways that the data get combined with other data sources and how those are used. They need to build and maintain systems that support this technological process but also can handle the disposition of data based on criteria beyond retention schedules. These systems and practices will need to be sensitive to the inter-dependencies that unexpected uses of the data find themselves.

A government organization releasing open data needs to work with tools that support checking the integrity of the data but also support removal of sensitive data. The repositories will need to support migration to new formats and hardware with a keen

sensitivity to handling existing or former ways of access that third parties may have used to access the data. In some cases, APIs and well-designed access portals can support this.

Beyond these technical concerns, licencing is a maintenance concern as well. It's clear that licencing encumbered by government regulation can prohibit how people use the data but it is also a worry for those maintaining the data. Over a long period of time, data deposited without standard open licences risks preventing those managing it from migrating it to new formats, systems, or otherwise continuing to make it available to the public. In addition, managing a variety of licences for different datasets could become a complex task in and of itself if many different and sometimes incompatible licences accompany the various datasets.

Standards

In general, open data repositories must address a number of standards problems, which is also true for open government data. The Open Data Standards Inventory project scanned many governments' open data repositories to identify file types in use and found that of the top 40 file formats, PDF was the most common for datasets⁷. This reveals part of the problem that digital curation faces with respect to open data. The PDF files, in this example, may serve some preservation purposes⁸ but they fail to fully support the Open Definition because they cannot be modified; and in some cases they discriminate against users who don't use software that supports proprietary formats.

As the Open Data Handbook explained, interoperability is a prime reason for open data. File formats that prohibit modification and make access problematic, run contrary to promoting interoperability. In other words an open government data repository cannot succeed in promoting the common goals of greater citizen participation, increased economic innovation, etc. so long as it is filled with proprietary or non-standard file formats.

From a digital curation stance aimed at facilitating successful open government data repositories, it would be wise to work with government organizations to obtain the source data for the repository in common, open standard formats such as CSV or ODF. Working with open file formats also precludes the dangers of proprietary formats. In the case of using proprietary formats like Microsoft's XLS format, a single vendor controls the format; no matter how ubiquitous its use may be, this is a big risk for the sake of long-term preservation, accessibility, and interoperability (the vendor may cease to support the format or stop existing altogether). Furthermore, it discriminates against users that either do not have the right systems or funds to purchase and use that vendor's software.

The interoperability issue is also important because the metadata used in various repositories' catalogues of their datasets varies from system to system⁹. To combine or use parts of different datasets together for new projects and insights, there needs to be common elements describing the data. The Open Data Standards Inventory

[<http://www.opengovpartnership.org/groups/opendata/resources>] suggested that governments collaborate on controlled vocabularies that could be used between different catalogues. Even where the metadata elements aren't exactly the same, they could be better identified through the consistent terminology of the controlled vocabulary.

Suggestions

Digital curators ought to involve themselves with the design and policies of the repositories and associated systems that enable access to open government data. They should make sure the Open Definition and accompanying best-practice principles of open government data are assured through the implementation and ongoing maintenance of the repository. This means strategically and collaboratively setting policy to require open data standard file formats with the most permissive licencing possible as well as a determination of the metadata standards to use.

With respect to metadata standards, while inter-governmental collaboration and compliance on controlled vocabularies is a laudable and worthy goal, there also ought to be systems operated by independent, non-proprietary third parties that serve as interchange hubs for translating the various metadata standards. This type of service has been set up and operated in commerce spaces for a long time, it may be a useful option for increasing the ongoing interoperability of open government data.

Conclusion

It is an impressive characteristic of open data, that so many organizations agree on its most fundamental qualities: the freedoms to use, modify, and share it for any purpose regardless of who you are.

A great deal of work has been completed to guide government organizations in how to properly make data open as well as to heed pitfalls. Those responsible for this data however, have a lot of work ahead of them as more government organizations adopt open data practices. They will need to manage the data and its accessibility with the entire range of preservation and curation requirements. They will also need to be sensitive to regulatory requirements (licencing, privacy, retention, etc.) and technical characteristics that are intimately joined with public usage (open standard file formats, interoperability, ongoing access through things such as API dependencies).

References

“The Annotated 8 Principles of Open Government Data.” [OpenGovData.org](http://opengovdata.org). 2015. Joshua Tauberer. <http://opengovdata.org> accessed 29 March 2015.

“About OpenDocument Format.” OpenDoc Society. <http://opendocumentformat.org/aboutODF/> accessed 2 April 2015.

Cabinet Office. 18 June 2013. *G8 Open Data Charter*. UK Presidency of G8 2013 Accessed 2 April 2015 from <https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex>

Canada. *Canada's Action Plan on Open Government*. Ottawa, Ont: Govt. of Canada, 2012. <<http://www.deslibris.ca/ID/232301>>. <http://open.canada.ca/en/canadas-action-plan-open-government>

Canada. *Digital Canada 150*. 2014. <http://epe.lac-bac.gc.ca/100/201/301/weekly_checklist/2014/internet/w14-14-U-E.html/collections/collection_2014/ic/Iu64-48-2014-eng.pdf>.

Canada, and Pierre-Luc Dusseault. *Open Data: The Way of the Future : Report of the Standing Committee on Government Operations and Estimates*. 2014. <http://epe.lac-bac.gc.ca/100/201/301/weekly_checklist/2014/internet/w14-26-U-E.html/collections/collection_2014/parl/xc70-1/XC70-1-1-412-5-eng.pdf>

Dodds, L. and Newman, A. 2015. *Open Data Maturity Model*. Open Data Institute. <http://theodi.org/guides/maturity-model> accessed 2 April 2015.

European Commission. 24 July 2014. *Guidelines on recommended standard licences, datasets and charging for the re-use of documents*. Accessed 2 April 2015 from ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6421

McKinney, J., Guidoin, S. and Marczak, P., Open Data Working Group. 9 March 2015. *Open Data Standards Inventory*. Open Government Partnership. Access on 2 April 2015 from <http://www.opengovpartnership.org/groups/opendata/resources>

McLeod, Julie. 2012. "Thoughts on the opportunities for records professionals of the open access, open data agenda". *Records Management Journal*. 22, no. 2: 92-97.

“The Open Definition.” [Open Definition](http://opendefinition.org). Open Knowledge Foundation. <http://opendefinition.org> accessed 29 March 2015.

Organisation for Economic Co-operation and Development. *Denmark, Efficient E-Government for Smarter Public Service Delivery*. Paris: OECD, 2010.
<<http://dx.doi.org/10.1787/9789264087118-en>>.

Park, Eun G, and Sam Oh. 2012. Examining Attributes of Open Standard File Formats for Long-term Preservation and Open Access. *Information Technology and Libraries* 31, no. 4: 46-67. <http://ejournals.bc.edu/ojs/index.php/ital/article/view/1946>.

Veljkovic N., Bogdanovic-Dinic S., and Stoimenov L. 2014. "Benchmarking Open Government: An Open Data Perspective". *Government Information Quarterly*. 31, no. 2: 278-290.

"Welcome to Open Government Data." [Open Government Data](http://opengovernmentdata.org). Open Knowledge Foundation. <http://opengovernmentdata.org> accessed 29 March 2015.

"What is Open Data?" [Open Data Handbook](http://opendatahandbook.org/en/what-is-open-data). Open Knowledge Foundation. <http://opendatahandbook.org/en/what-is-open-data> accessed 29 March 2015.

Zuiderwijk A., Janssen M., and Davis C. 2014. "Innovation with Open Data: Essential Elements of Open Data Ecosystems". *Information Polity*. 19, no. 1-2: 17-33.

"Open Data Policy Guidelines." [Sunlight Foundation](http://sunlightfoundation.com/opendataguidelines). Sunlight Foundation. <http://sunlightfoundation.com/opendataguidelines> accessed 29 March 2015.

- 1 Dodds, L. and Newman, A. 2015. Open Data Maturity Model. Open Data Institute.
<http://theodi.org/guides/maturity-model> accessed 2 April 2015. p. 3
- 2 Canada. Canada's Action Plan on Open Government. Ottawa, Ont: Govt. of Canada, 2012.
<<http://www.deslibris.ca/ID/232301>>. <http://open.canada.ca/en/canadas-action-plan-open-government>
- 3 Cabinet Office. 18 June 2013. G8 Open Data Charter. UK Presidency of G8 2013 Accessed 2 April 2015 from <https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex> p. 1
- 4 McKinney, J., Guidoin, S. and Marczak, P., Open Data Working Group. 9 March 2015. Open Data Standards Inventory. Open Government Partnership. Access on 2 April 2015 from <http://www.opengovpartnership.org/groups/opendata/resources>
- 5 Zuiderwijk A., Janssen M., and Davis C. 2014. "Innovation with Open Data: Essential Elements of Open Data Ecosystems". Information Polity. 19, no. 1-2: 17-33. p. 22
- 6 McLeod, Julie. 2012. "Thoughts on the opportunities for records professionals of the open access, open data agenda". Records Management Journal. 22, no. 2: 92-97.
- 7 McKinney, J., Guidoin, S. and Marczak, P., Open Data Working Group. 9 March 2015. Open Data Standards Inventory. Open Government Partnership. Access on 2 April 2015 from <http://www.opengovpartnership.org/groups/opendata/resources> p. 17
- 8 Park, Eun G, and Sam Oh. 2012. Examining Attributes of Open Standard File Formats for Long-term Preservation and Open Access. Information Technology and Libraries 31, no. 4: 46-67. <http://ejournals.bc.edu/ojs/index.php/ital/article/view/1946>. p. 45
- 9 McKinney, J., Guidoin, S. and Marczak, P., Open Data Working Group. 9 March 2015. Open Data Standards Inventory. Open Government Partnership. Access on 2 April 2015 from <http://www.opengovpartnership.org/groups/opendata/resources> p. 14