Welcoming a New Era of Dynamic Open Data

by Mike Bergman - Monday, January 27, 2014

http://www.mkbergman.com/1707/welcoming-a-new-era-of-dynamic-open-data/

It's Time to Move Beyond Static Dataset Dumps

It would be an understatement to say that open data has been transforming how government does business. Over the past five years -- ranging from national governments such as the United States and the United Kingdom to hundreds of local governments and municipalities and all forms of government in between -- a veritable revolution in opening up data to the public has been underway. The open data in government (OGD) movement has spawned an entirely new cottage industry in open data advocacy and tools. Literally hundreds of government organizations are committed to open data, supported by an ecosystem of advocacy, technology and consulting groups.

Open data, of course, is not limited to governments. Open data in science and from the Web and for-profit entities are legitimate focal points in their own right. But, because data generated by governments are both sanctioned and developed using taxpayer monies, open data in government (OGD) occupies a special place in the conversation. Now, with experience and practice, we are beginning to see a generational shift in how open data is being handled by governments. The first generation, still mostly the current practice, was built around the idea of just making the data public and open. This current generation of open data is characterized by the publishing of datasets via catalogs. The datasets are static, unconnected and dumb. Mostly, too, the data within those datasets are poorly described and documented, often lacking standard metadata. What is now exciting, however, is the emergence of what can best be called *dynamic open data*. What this is and how it offers advantages is the focus of this article.

The 8 Initial Principals of Open Government Data

In October 2007, 30 open government advocates met in Sebastopol, California to discuss how government could open up electronically-stored government data for public use. Up until that point, the federal and state governments had made some data available to the public, usually inconsistently and incompletely, which had whetted the advocates' appetites for more and better data. The conference, led by Carl Malamud and Tim O'Reilly and funded by a grant from the Sunlight Foundation, resulted in eight principles that, if implemented, would empower the public's use of government-held data. These

principles, no longer online, were summarized by Joshua Tauberer in his Open Government Data book as:

1. Data Must be **Complete**
   All public data are made available. Data are electronically stored information or recordings, including but not limited to documents, databases, transcripts, and audio/visual recordings. Public data are data that are not subject to valid privacy, security or privilege limitations, as governed by other statutes.
2. Data Must be **Primary**
   Data are published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.
3. Data Must be **Timely**
   Data are made available as quickly as necessary to preserve the value of the data.
4. Data Must be **Accessible**
   Data are available to the widest range of users for the widest range of purposes.
5. Data Must be **Machine Processable**
   Data are reasonably structured to allow automated processing of it.
6. Access Must be **Non-Discriminatory**
   Data are available to anyone, with no requirement of registration.
7. Data Formats Must be **Non-Proprietary**
   Data are available in a format over which no entity has exclusive control.
8. Data Must be **License-free**
   Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed as governed by other statutes.

These basic principles were then updated and re-phrased by the Sunlight Foundation in August 2010 to now number 10 principles, including the use of open standards, making data permanent, and keeping usage costs to an absolute minimum. All of these are laudable points. Each may or may not be provided in a fully open way by any given governmental entity.

This first step in the open data process has led to systems that are oriented to posting and publishing downloadable datasets. Existing open government data platforms, for example, such as Socrata or DKAN, can best be described as catalog systems. Listings of datasets with associated descriptions and metadata are presented. Users or the public may then chose among one or more machine-readable formats to download the entire dataset.

The 5 Added Principles of Dynamic Open Data

Of course, simply throwing data over the fence does not make it useful. Once we can get past the first threshold of making data publicly accessible, we next face the challenge of making that data meaningful and relevant. Since relevance is in the eye of the user, we no longer can think about information solely in terms of static, dumb datasets. We now need to expose the underlying data dynamically, such that users may request and filter and correlate what they need and only what they need.

Thus, there are five principles -- or dimensions -- by which we need to judge next-generation dynamic open data:

1. Data Should be **Filterable**

   Data should be selectable by type (class), attribute or value such that only the data of interest is exposed to the user. This means the data should be structured in some way with facets that can be used dynamically to filter and make those selections.

2. Data Should be **Atomic**

   Data should be exposed as individual entities or concepts with their attributes and values. The unit of manipulation thus becomes the datum, rather than the dataset.

3. Data Should be **Connected**

   Because we are now collecting by datum and not dataset, connections between relevant things must be made explicit across relevant datasets. Similar things should be retrievable together. To achieve this aim, some schema or data definition framework must be layered over the data and datasets.

4. Data Should be **Expandable**

   Since new data and new instances and new datasets will constantly arise, the design of the overall data management system must itself be "open", enabling expansion of the available datastore at acceptable cost and effort.

5. Data Should be **Documented**

   In order for these dynamic selections to be achievable, the data in the system must be fully documented, specifically including the full description and units used for attributes and values and the scope of entities and concepts. Only through such complete documentation can accurate connections and relevant selections per above be made.

There is no set order to the principles above. They are presented in the order shown so as to help remember them through the **FACED** mneumonic.

Parallels with Linked Data

Though the principles above do not call out [linked data](#) as a requirement, they do share many parallels with the early growth and maturation of linked data. A number of years back Fred Giasson and I commented on [When Linked Data Rules Fail](#). Two of the points made in that article are the absence of suitable data descriptions and lack or wrong connections in data.gov and the NY Times datasets. I subsequently expanded on these types of problems in [Practical P-P-P-Problems with Linked Data](#).

Official data from governments can avoid many of the provenance issues associated with general linked data, but in other areas there are important parallels. Like any emerging new practice, it takes a while to learn and formalize best practices. It is not surprising that we are seeing open data in government needing to transition from dumb datasets to actionable information. Making data actionable is when government information assets will finally become effective for the broader public.

Also, like linked data, it is likely the platforms built around semantic technologies and knowledge graphs (schema) will also come to the fore. Our own [Open Semantic Framework](#) is one such example, but there are a few now emerging in the linked data and semantic technology space. It will be through different practices and these newer platforms that we will see the next generation of open government data truly emerge.

PDF generated by *AI3:::Adaptive Information* blog