The Continuing Evolution of Data.gov

Dennis D. McDonald, Ph.D.

The Continuing Evolution of Data.gov
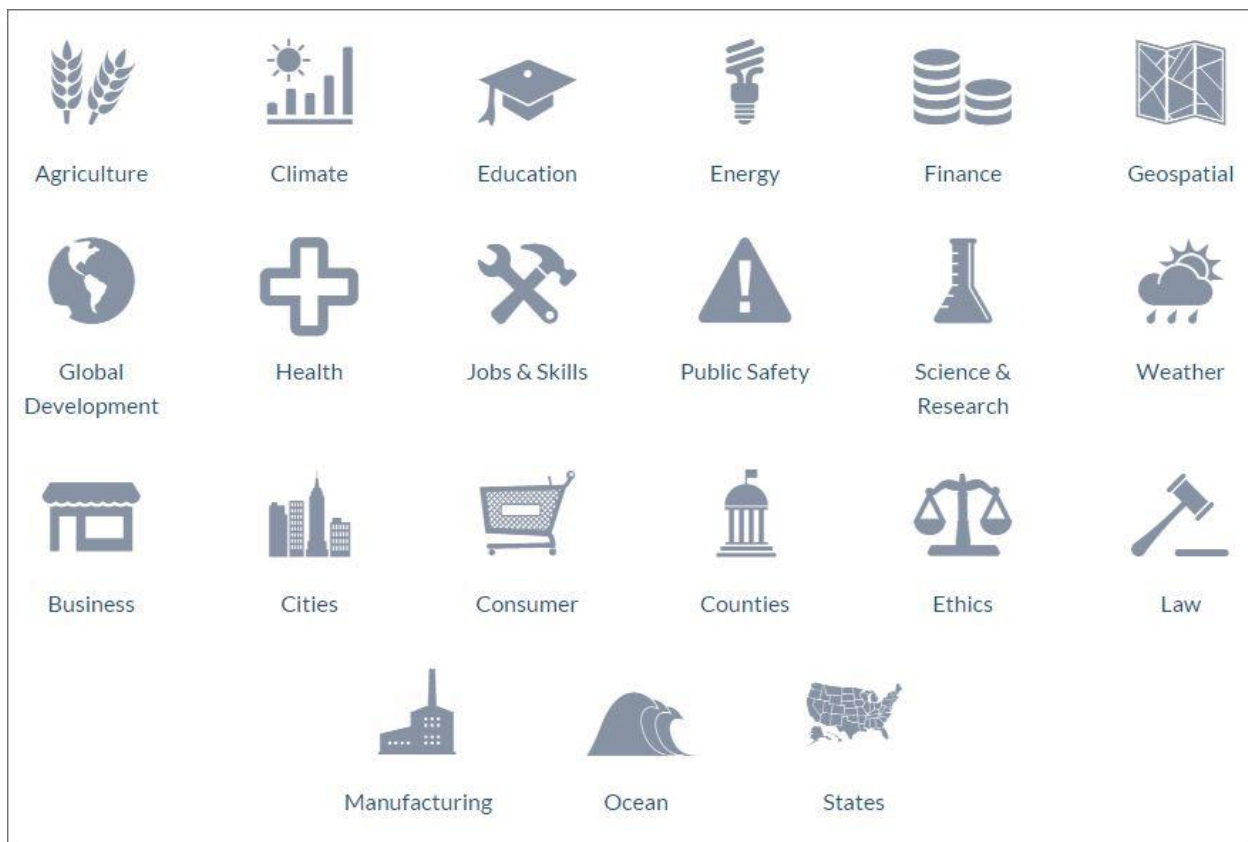
By Dennis D. McDonald, Ph.D.[1]

November 20, 2014

Introduction

Last night's Open Data Leaders meetup in Washington DC provided an informative view of the inner workings of Data.gov which calls itself "The home of the U.S. government's open data."

Started in 2009 with an initial 47 datasets, the site now provides an online catalog to over 132,445 datasets. These are spread across multiple disciplines are shown below:

What I learned

I learned several things while listening to Data.gov's Philip Ashlock, Hyon Kim, and Rebecca Williams:

- The site really took off in 2013 when the Obama Administration began requiring Federal agencies to inventory and supply metadata describing data sets in a specified format.
- The site provides a front-end catalog of metadata that supports searching across data sets provided by multiple agencies. It is the responsibility of the agencies themselves to provide the actual data.
- Data.gov makes extensive use of open source tools including Github (source code management and issue tracker), CKAN (data and API catalogs), and WordPress (web portal).
- Data.gov, housed in the General Services Administration (GSA), is an active participant in Project Open Data which is run by the White House's Office of Management and Budget (OMB) and Office of Science and Technology Policy (OSTP).
- While the contents and quality of individual datasets is the responsibility of the providing agency, Data.gov staff are working with OMB on the development of qualitative and quantitative metrics to describe dataset contents.
- As evidence of how seriously this Administration takes "open data," OMB has promoted "open data" as one of its "cross agency priority goals." This is how OMB describes the Open Data "Goal Statement": *Fuel entrepreneurship and innovation and improve government efficiency and effectiveness by unlocking the value of government data and adopting management approaches that promote interoperability and openness of this data.*
- A handful of U.S. states currently supply dataset metadata to Data.gov in the prescribed format, thereby making their catalogs searchable along with Federal data.
- Data.gov has not looked at inferring taxonomies to expand searches in the background but has instead focused on structured metadata to support the filtering process during a search.
- Occasionally questions arise during search that can best be handled by the data supplier – someone at the data-supplying agency itself. One possible approach to handling such user support issues might be to incorporate contact names of agency staff as part of retrieved results on Data.gov.

Different Approaches

During the question and answer period I compared the two stage approach being taken by Data.gov with the approach being developed by NOAA in its "Big Data Partnership" program.

The two stages in Data.gov are the operation of a unified catalog of metadata that provides access to multiple data sources (maintained by individual agencies). The advantage of this approach is that the individual agencies understand their data the best and are responsible for the processes and systems that generate the data in relation to their program objectives. Data.gov provides a unified "front end" via its search engine and requests for data are handed over to agencies for fulfillment.

While this division of labor makes sense the downside, suggested earlier, is in terms of customer support. When a search "fails" whose fault is it? What if the inquiry is subject-specific and requires intervention by a subject matter expert? Plus, what if the data themselves are at fault which was mentioned by one of the questioners at the meeting; what responsibility does Data.gov have for that?

The program being designed by NOAA also has two stages. NOAA provides public access to data that are "inherently governmental" and tied directly to its carrying out its legislated program objectives. Yet NOAA also generates a substantial amount of data that might also be useful to the public were it available for exploitation. Development of such a program whereby private sector organizations agree to provide cloud based access for such data at cost is currently ongoing.

These are different approaches but it is the essence of the concept of open data that the data are made available for exploitation and re-use. It would be wrong to suppose that all approaches to making open data useful have already been tried so I don't see these two approaches being taken by the Federal government to be contradictory. Now is the time to experiment.

Implications

During the question and answer period the Data.gov team was asked what it was like to work with the different agencies to get them to provide metadata in the specified format. Those who have worked with open data and enterprise data programs will recognize the response:

- You need a technically competent person at the source agency to work with.
- Larger agencies with more structured IT operations are better able to supply the appropriate expertise.
- Smaller agencies with less resources require more handholding – but the flip side of this is that they have less history and culture issues to overcome.

Keep in mind that these efforts are being managed by GSA and OMB. These are "general purpose" agencies whose managerial and support services are designed for the sake of efficiency to support multiple agency missions. Sometimes there is friction between OMB and individual agencies if OMB requirements are perceived to be indirectly related to the agency's programs or if the actions required by OMB are not explicitly funded in the agency's budget.

Conclusions

It is a testament to this Administration's emphasis on "open data" that we have gotten this far with efforts such as Data.gov. Just as important is the manner in which this effort has been managed with a large component of shared energy among technically inclined techies and data wranglers in many different organizations. While it is true that a platform such as GitHub is not really designed to be as user friendly as, say, Facebook, the fact is that the sharing of technical expertise among mid-level IT staff and data administrators in different governmental agencies has probably been at least as important to open data progress as the Administration's top down support.

Related reading:

- *[Justifying Collaboration in Complex Programs such as Federal Acquisitions](#)*
- *[Why best practices won't fix federal IT](#)*
- *[Challenges in Turning Federal Digital Strategy into Mobile Reality](#)*
- *[Scoping Out the 'Total Cost of Standardization' in Federal Financial Reporting](#)*
- *[How Should DATA Act Implementation Impact Federal Project Management Practices?](#)*
- *[A Project Manager's Perspective on the GAO's Federal Data Transparency Report](#)*
- *[Using Collaboration Technologies to Accelerate Innovation in Federally Funded R&D Programs](#)*