Making Sense of Open Government Data

Li Ding
Tetherless World Constellation,
Rensselaer Polytechnic Institute
110 8th St. Troy NY12180, USA
dingl@cs.rpi.edu

James R. Michaelis
Tetherless World Constellation,
Rensselaer Polytechnic Institute
110 8th St. Troy NY12180, USA
michaj6@cs.rpi.edu

Deborah L. McGuinness
Tetherless World Constellation,
Rensselaer Polytechnic Institute
110 8th St. Troy NY12180, USA
dlm@cs.rpi.edu

Jim Hendler
Tetherless World Constellation,
Rensselaer Polytechnic Institute
110 8th St. Troy NY12180, USA
hendler@cs.rpi.edu

ABSTRACT

Data.gov, a major distributor of raw US government data, has published thousands of raw datasets on the Web for public access. While these datasets provide useful information, their potential has not yet been fully realized due at least partially to a number of usability-related issues. In this work, we investigate ways to make sense of existing open government data using semantic web technologies. We also demonstrate strategies for turning open government data into linked government data and present several case studies to illustrate the role of linked government data in making sense of government data.

Keywords

Linked Data, Linked Government Data, Data.gov

1. INTRODUCTION

Availability of US government data on the Web is steadily increasing. Recently, thousands of raw datasets have been published via Data.gov, the main distributor of US government data. Although opening up government data on the Web has helped promote government transparency, these efforts have so far not guaranteed easily consumable data. With online tools and data access interfaces, users can access and browse government datasets. However, it is still very difficult for users to integrate distributed government data to conduct unexpected studies.

Currently, a major issue with open government data is to help users to make sense of government data. Some datasets are incomplete and need additional information, some use special codes and acronyms, and some lack metadata on the datatypes and meaning of data entries. Meanwhile, Data.gov is trying to help users address these issues: metadata of datasets are often provided through documents in PDF format with hundreds of pages. In this work, we show paths toward meaningful government data and explain how linked government data can be of benefit via several case studies collected from the Data-gov project.

1.1 Open Government Data at US

The Presidential Open Government Directive [1] demands that US government agencies publish government data online with the following guidelines:

- *Publish Government Information Online*
- *Improve the Quality of Government Information*
- *Create and Institutionalize a Culture of Open Government*
- *Create an Enabling Policy Framework for Open Government*

Following this direction, increasing amounts of government data are published online via Data.gov, including 1,276 raw data catalogs, 494 tool catalogs, and 167,966 geodata catalogs as of March 29, 2010. While these datasets provide useful information, their usability (through applications and direct data access) is often limited. Some of the most interesting datasets from Data.gov are from the raw data catalogs, because they are meaningful instead of being merely geoshapes. Additionally, they expose full access to data instead of hiding data from the visual data access interface. In the rest of this paper, we focus on the raw data catalogs.

1.2 Linked Government Data

The Semantic Web offers a promising infrastructure for managing US government datasets. Following the principles of Linked Data [2], US government data can be published on the Semantic Web in the form of Linked Government Data (LGD) [3] using semantic web technologies such as RDF, RDFS, SPARQL and RDFa. LGD shows great potential in making government data more accessible and useful, as demonstrated, for example, by efforts in the UK [4] and the US [5].

1.3 The Data-gov Project

The Tetherless World Constellation at RPI is producing LGD derived from US government data published by Data.gov and other sources through the Data-gov Wiki[1]. Our work on the Data-gov project includes LGD generation, LGD-based demo creation, and LGD education. So far, we have published 5 billion LGD triples based on hundreds of datasets from Data.gov and other government-related sources.

One of our ultimate objectives is to improve the user experience in using US government data through the use of LGD and LGD based-applications. This objective not only demands convenient data access but also requires designs that can make it easy for users to correctly understand the meaning of government data.

[1] http://data-gov.tw.rpi.edu

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

2. ENRICHING MEANING USING LGD

Figure 1 illustrates several ways in which government dataset usability can be improved using LGD:

- First, not all government data are self-contained, and some are published as a collection of interrelated complementary datasets. In order to making sense of one such dataset, the related dataset should also be loaded. Here, if we can explicitly declare the inter-connections among related datasets, users can avoid run-time alignment computations in data integration. Moreover, if adopting LGD publishing principles (e.g. using dereferencable URIs for concepts defined externally), agents can directly surf the Web of data without asking users.
- Second, we should also carefully curate links from a dataset to its metadata. When dealing with data tables from government datasets, both headers and values may be hard to understand due to their use of acronyms and codes. By linking to the metadata, which provides interpretations for the elements in data that are hard to understand, users can not only understand the content of datasets but they also can perform dataset validation tasks, such as range checking.
- Third, Data.gov datasets in different domains may be correlated by referencing the same concept (such as common fiscal years and states). We can therefore link such correlated data to support data analysis on government data (e.g. covariance analysis).

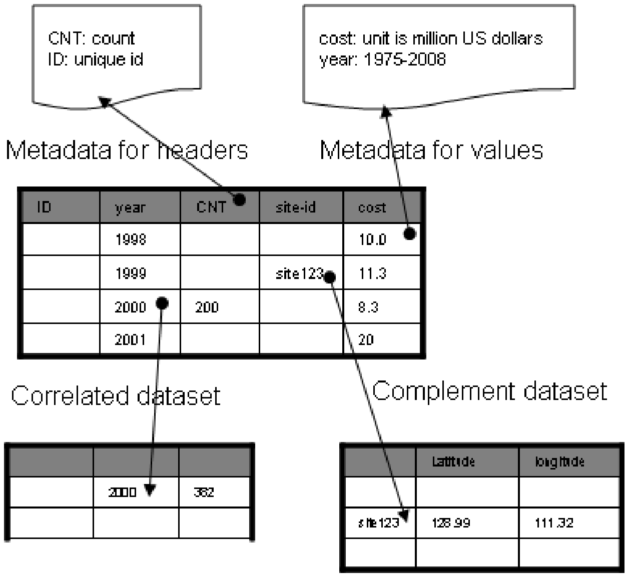


Figure 1. Links to auxiliary data supporting better understanding of government datasets.

In what follows, we present several case studies, each based on an LGD application developed at RPI, to illustrate the role of LGD in making sense of government data and thus enhancing the impact and utility of open government applications.

2.1 Case Study 1: Complimentary Datasets

Upon studying the environmental statistics of the US, we found an interesting Data.gov dataset titled "Clean Air Status and Trends Network (CASTNET): Ozone" and published by the U.S. Environmental Protection Agency (EPA). The dataset records daily ozone readings at different monitoring sites (identified by site id), but lacks information about the geographic location of the sites. In order to provide a map to intuitively show the distribution of Ozone reading throughout the U.S., we search on the Web and found a complementary dataset (containing geo-coordinate location information for every monitoring site) from epa.gov. Aligned by site-id, the two datasets complement each other and yield Figure 2: the former dataset helps determine the size of circles in the map based on average ozone readings at the site, and the latter enables plotting of monitor sites on a map.

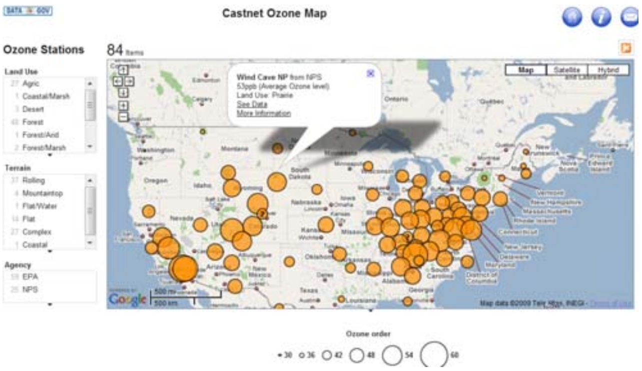
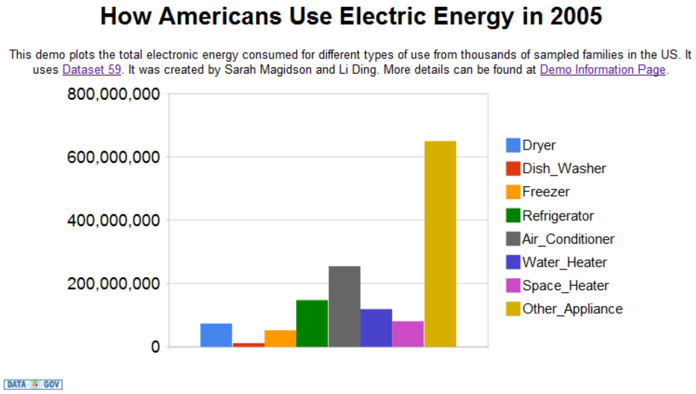


Figure 2. ozone map demo links ozone reading and location

Demo: http://data-gov.tw.rpi.edu/wiki/Demo:_Castnet_Ozone_Map
Dataset(from data.gov): http://data-gov.tw.rpi.edu/wiki/Dataset_8
Dataset(from epa.gov): http://data-gov.tw.rpi.edu/wiki/Dataset_10001

2.2 Case Study 2: From Data to Metadata

A meaningful plot of government data demands an understanding of metadata. Now let's look into a dataset from the Department of Energy (DOE) called "Residential Energy Consumption Survey (RECS) Files, Energy Consumption, 2005". In this dataset, a majority of properties are given in acronym form – requiring the use of external references to explain their meanings. For example, the property "BTUELRFG" means "Electric Refrigerator Use (Estimated)", with corresponding values defined as "USE IN THOUSANDS OF BTU (1 - 9999998) 9999999 NOT APPLICABLE". By linking the dataset to metadata (in PDF), we built a demo with meaningful labels and values (see Figure 3).



This demo plots the total electronic energy consumed for different types of use from thousands of sampled families in the US. It uses Dataset 59. It was created by Sarah Magidson and Li Ding. More details can be found at Demo Information Page.

Figure 3. Energy consumption table with correct labels and values

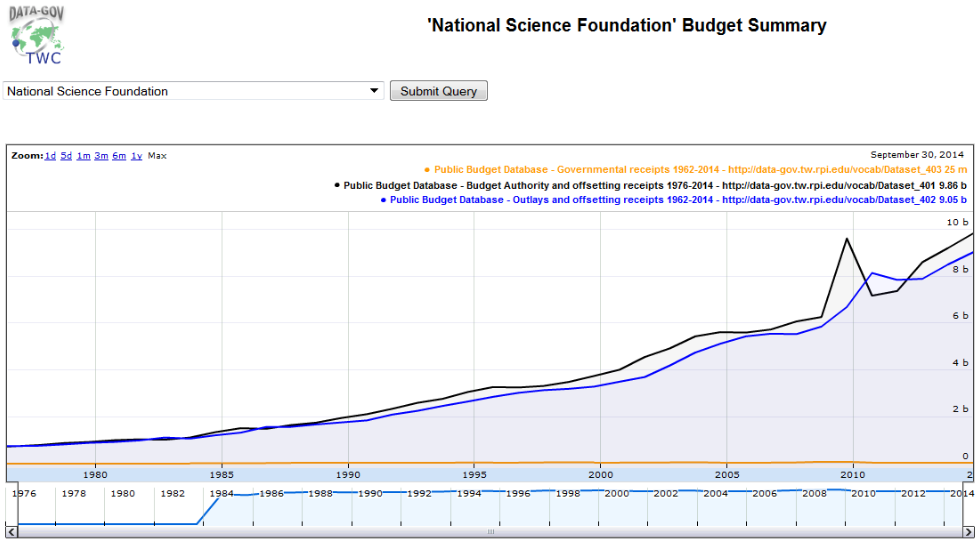
Demo: http://data-gov.tw.rpi.edu/wiki/Demo:_Electric_Energy_Consumption
Dataset(from data.gov): http://data-gov.tw.rpi.edu/wiki/Dataset_59
Data dictionary(p): http://www.eia.doe.gov/emeu/recs/recspubuse05/layoutfiles/RECS05layoutFILE11.csv
Data dictionary (value): http://www.eia.doe.gov/emeu/recs/recspubuse05/pdf/recs05codebook.pdf

2.3 Case Study 3: Related Data

It is useful to display correlated datasets side-by-side to run cross-validation as well as knowledge discovery. The meaning of a single dataset can be explained or enriched both showing related datasets. This use case focuses on three correlated datasets from the Office of Management and Budget (OMB): “Budget Authority and offsetting receipts 1976-2014”, “Outlays and offsetting receipts 1962-2014” and “Governmental receipts 1962-2014”. In Figure 4, the three datasets are interlinked through a common set of properties, enabling dataset cross-validation. For example, the property “bureau_name” can be used to check the budget information about a selected agency across all three datasets. However, the value “National Water Commission” is known to only occur in the second dataset – indicating the possibility of flaws in one or more datasets.

2.4 Case Study 4: Meaningful Visualization

A correct value from a dataset may not be enough to yield a meaningful demo. For example, we should compare state-wise statistics after normalizing by state population to avoid bias. This use case is based on a dataset from the Institute of Museum and Library Services called the “State Library Agency Survey: Fiscal Year 2006”. In this dataset, both book volume and local population figures have been provided. In Figure 5, the left-hand side demo only used the total book volume which could be biased by state population. Therefore, the second version of the same demo (right-hand side) compares average book volume per person in different states. A contrast between the older version and new version of the demo can be found on California and New York - they are less knowledgeable after taking the population based normalization.

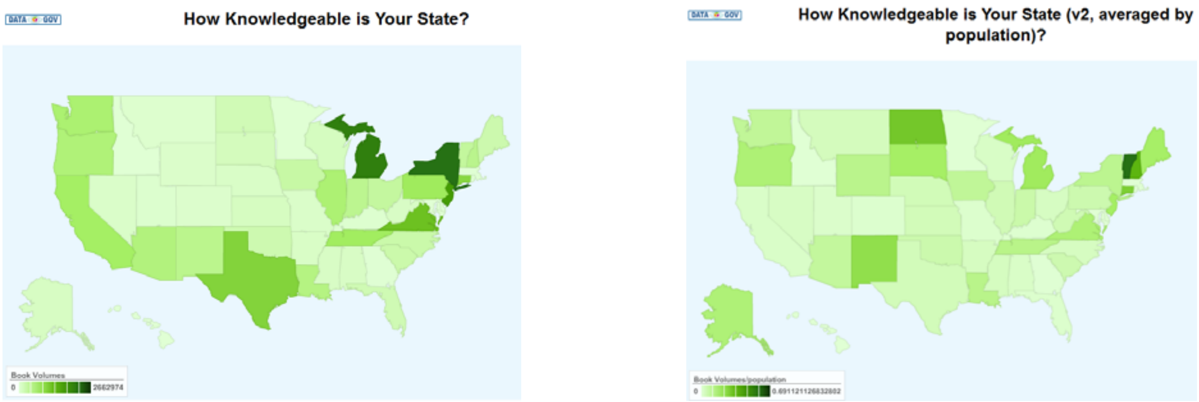


'National Science Foundation' Budget Summary

National Science Foundation ▾ Submit Query

Figure 4. Budget Authority, Outlays and receipts are published table with correct labels and values

Demo: http://data-gov.tw.rpi.edu/wiki/Demo:_Agency_Budget_Summary
Dataset: http://data-gov.tw.rpi.edu/wiki/Dataset_401 (Budget Authority and offsetting receipts 1976-2014)
Dataset: http://data-gov.tw.rpi.edu/wiki/Dataset_402 (Outlays and offsetting receipts 1962-2014)
Dataset: http://data-gov.tw.rpi.edu/wiki/Dataset_403 (Governmental receipts 1962-2014)



DATA GOV
How Knowledgeable is Your State?
Book Volumes
0
2662974
DATA GOV
How Knowledgeable is Your State (v2, averaged by population)?
Book Volumes/population
0
0.691121126832802

Figure5. “how knowledgeable is your state” demo got more meaningful with context data

Demo: http://data-gov.tw.rpi.edu/wiki/Demo:_How_Knowledgeable_is_Your_State
Dataset: http://data-gov.tw.rpi.edu/wiki/Dataset_353 (State Library Agency Survey: Fiscal Year 2006)

3. Conclusion

Based on the above case studies, we have seen the needs and benefits of LGD in making sense of government data and related applications. It should be noted that currently, the creation of LGD involves non-trivial human interaction. For instance, the second EPA dataset in Case Study 1 was manually found using Google search. In Case Study 4, the library data should be linked to population data to yield less biased results.

The creation of LGD, especially high quality LGD, certainly needs non-trivial human user contributions. To address this, we are working on a social web based platform to enable crowdsourcing. Ultimately, the creation of LGD will need "social data networks" with machines and communities of citizens interacting with government data to jointly solve the challenges.

4. ACKNOWLEDGMENTS

This work is funded in part by a grant from DARPA's Transformational Convergence Technology Office.

5. REFERENCES

[1] Open Government Directive, 2009, http://www.whitehouse.gov/open/documents/open-government-directive .

[2] Berners-Lee, T. *Putting government data online*. 2009. http://www.w3.org/DesignIssues/GovData.html .

[3] Berners-Lee, T. *Linked data*. 2007. http://www.w3.org/DesignIssues/LinkedData.html.

[4] Alani, H.; Dupplaw, D.; Sheridan, J.; O'Hara, K.; Darlington, J.; Shadbolt, N.; and Tullo, C. *Unlocking the potential of public sector information with semantic web technology*. In ISWC/ASWC, 708–721. 2007.

[5] Ding,L.; DiFranzo,D.; Graves,A.; Michaelis,J.; Li,X.; McGuinness,D.; and Hendler,J., *Data-gov Wiki: Towards Linking Government Data*. In Proceedings of the AAAI Spring Symposium on Linked Data Meets Artificial Intelligence. 2010.