

The Data-gov Wiki: A Semantic Web Portal for Linked Government Data

Li Ding, Dominic DiFranzo, Deborah L. McGuinness, Jim Hendler

Tetherless World Constellation, RPI
110 8th St, Troy, NY12180, USA
{dingl, difrad, dlm, hendler}@cs.rpi.edu

Sarah Magidson

University of Chicago
5801 South Ellis Avenue, Chicago, IL 60637
magidson@uchicago.edu

ABSTRACT

The Data-gov Wiki is the delivery site for a project where we investigate the role of linked data in producing, processing and utilizing the government datasets found on data.gov. The project has generated over 2 billion triples from government data and a few interesting applications covering data access, visualization, integration, linking and analysis.

1. INTRODUCTION

The recently launched data.gov website has released 700+structured government datasets from 44 different US government agencies for public access. Our Data-gov Wiki (<http://data-gov.tw.rpi.edu/>) is the delivery site for a project where we investigate the role of linked data [1] in producing, processing and utilizing the government datasets found in data.gov. The wiki hosts a number of demos/tools and describes the techniques used for making data.gov available using the Web standards developed in the World Wide Web Consortium's Semantic Web Activity.

To realized linked government data (gov-data) [2,3], we exhibit the following semantic web based components via Data-gov Wiki: (i) convert gov-data into RDF for machine- friendly data access; (ii) enrich and link gov-data via various data processors (e.g. extraction, normalization, and mapping); and (iii) show the value of linked gov-data using interesting applications built on the low-hanging fruit of the semantic web (e.g. SPARQL engine) and the Web in general (e.g. Google Visualization API).

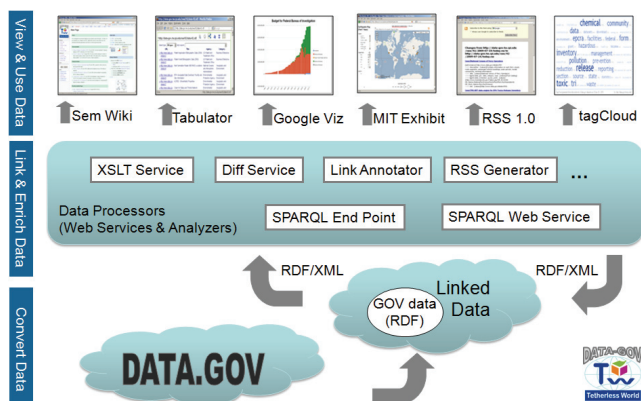


Figure 1 Semantic Web Architecture for Government Data

2. CONVERTING GOV-DATA

The datasets from data.gov are typically structured as tables/spreadsheets, and they are published in different formats such as CSV/TXT, Excel, XML, and KML. Currently, we have created 16 RDF datasets covering 187 datasets listed on data.gov (171 datasets from the EPA are subsets of three big datasets). The translated RDF datasets contribute a total of 2,927,398,352 triples involving 2,526 properties.

Our conversion process adopted the following principles: The conversion should be kept minimal, just enough to preserve the structure and content of gov-data, and the converted RDF data can always be enriched by advanced data processors in the future. The converted data should be accessible on the Web; therefore, RDF/XML was chosen for data publishing and all URIs in the converted data are dereferenceable via HTTP protocol. The ontology for the converted data should also be extensible, and we leverage Semantic MediaWiki (SMW) to collect Web users' edits. Provenance of the conversion should also be recorded, and we used popular ontologies including Dublin Core (DC) and FOAF.

We also observed some issues in reusing gov-data. For example, some datasets were encoded in hard-to-parse formats, some datasets were published via query-based interactive access points, and some datasets mentioned values that are meaningless without being further interpreted by other datasets. These issues justify the importance of linked data in gov-data reuse. For more details, see http://data-gov.tw.rpi.edu/wiki/Current_Issues_in_data.gov.

3. DEMOS AND TECHNOLOGY USED

With the converted RDF gov-data, we then built applications using existing (semantic) web tools. The applications were developed with two purposes: demonstrating the value of the linked gov-data and serving as examples for web developers to learn and adopt semantic web technologies. All applications are listed at <http://data-gov.tw.rpi.edu/wiki/Demos>.

1.1 Data catalog on the Data-gov Wiki

Dataset 92 (<http://www.data.gov/details/92>) provides metadata about the datasets listed on data.gov. Each dataset is described by 52 properties. Currently, data.gov provides faceted search (keywords, format, agency, and category) and browsing (100 items per page) for finding datasets. However, users cannot locate datasets using other properties (e.g. title or description), and users can only contribute ratings to the datasets. We leveraged SMW to add some new features to the Data-gov Wiki:

- Replicating the catalog metadata. We used JENA to convert the RDF version of Dataset 92 into MediaWiki's XML "wikidump" format and preserve the RDF data using SMW annotations on each dataset's wiki page.
- Integrating more metadata. The original metadata about a dataset were enriched by more metadata contributed by manual wiki editing (e.g. issues with a dataset) and computer programs (e.g. number of triples).
- Enabling customized data access. Users can access data via MediaWiki's keyword search and SMW's query, for example http://data-gov.tw.rpi.edu/wiki/Data.gov_Catalog.

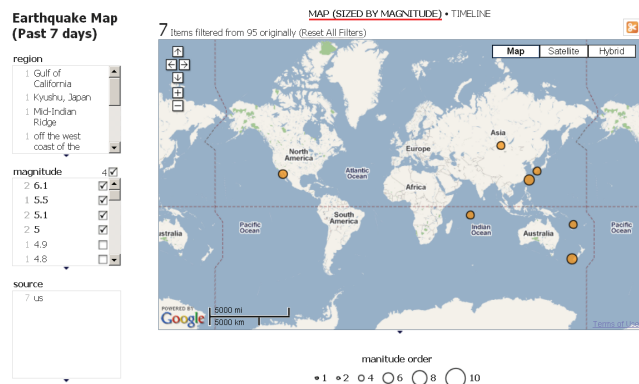
1.2 Data Visualizations

A number of (interactive) visualizations of the data from data.gov have been built to show (i) how to hook up the converted RDF data with conventional Web tools, and (ii) the value of linked

gov-data. We see a generic design for developing gov-data visualization applications following Jeni Tension's blog (<http://www.jenitension.com/blog/node/113>, July, 2009):

- Query Data. We execute a SPARQL query with a FROM clause pointing to the linked gov-data through a SPARQL query web service (<http://onto.rpi.edu/sw4j/sparql.html>) and get SPARQL/XML encoded query results.
- Convert Data. We then convert the query results to the specific data structure required by visualization tools in JSON format. For the Google visualization API, we used Jeni's XSL stylesheet with an online XSLT processor, and for MIT Exhibit, we wrote some PHP conversion code.
- Visualize Data. This is conventional Web developing work.

Figure 2 shows a map of earthquakes of magnitude greater than 1M in the past 7 days (dataset 34). The map display uses the faceted browsing functions provided by MIT Exhibit.



<http://data.gov.tw.rpi.edu/wiki/Demo: Interactive Faceted Browser for Earthquake Data>

Figure 2. Worldwide 1M+ Earthquake Map - Past 7 days

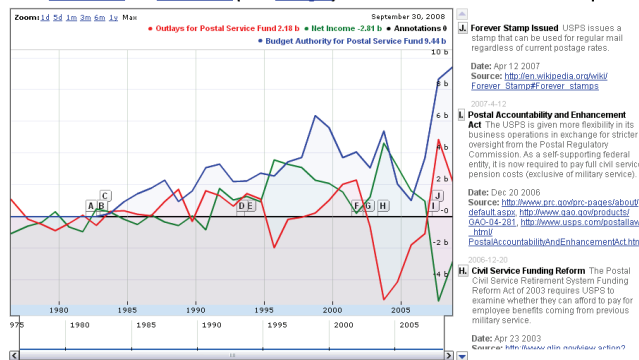
1.3 Data Integration and Linking

Figure 3 plots information about the same government account, the *Postal Service Fund*, from two different datasets, namely the federal budget authority (dataset 401), budget outlays (dataset 402), income and expense data collected from USPS website, and USPS related news collected from the Web. Examining this figure, interested readers might consider why there is a big drop in 2004 in budget outlays, and how that change can be linked to news released in the same year, e.g. the increases in postage rates or the changes in the USPS's business structure.

Tracking USPS Money:

Budget Authority, Outlays, Income, and Expenses

From Dataset 401 and Dataset 402 (from Data.gov) and a USPS data file on income and expenses



<http://data.gov.tw.rpi.edu/wiki/Demo: Timeline of USPS Money - Graph and Historical Events>

Figure 1 Budgets of Postal Service Fund (1962-2014)

1.4 Data Computations

Tracking the changes of the 700+ datasets listed on data.gov is not an easy job for human users. Unfortunately, data.gov does not offer an RSS feed for their datasets, and Twitterers' attempts to keep track (http://twitter.com/DataGov_Tweets) stalled on June 17, 2009. Therefore, we used semantic web tools to generate our own RSS (http://data.gov.tw.rpi.edu/wiki/RSS_Feeds) on a daily basis to keep track of (1) what datasets are available on data.gov and (2) which datasets have been recently added, updated or deleted (see figure 4). The former RSS feed was generated by a SPARQL query (converting vocabulary) and some tweaks on the generated RSS (sorting and placing channel descriptions to meet RSS readers). The latter RSS feed was derived from an online service (<http://onto.rpi.edu/sw4j/diff.html>) that computes and summarizes the difference between two RDF graphs at instance level, where each gov-data's metadata is stored as an instance.

Changes from <http://data.gov.tw.rpi.edu/raw/92/2009-07-19/today.rss> to <http://data.gov.tw.rpi.edu/raw/92/2009-07-24/today.rss>

[new]National Census of Ferry Operators

```
[add instance] http://www.data.gov/details/454
----> :description . [values]{Contains information on each ferry vessel,
route segments, passenger/vehicle boardings, peak periods, modal
connectivity, terminal information}
----> :title . [values]{National Census of Ferry Operators}
----> dgp92:data_gov_data_category_type . [values]{Tool Catalog}
----> dgp92:data_extraction_access_point . [values]
{http://www.transstats.bts.gov/Tables.asp?DB_ID=616&
DB_Name=National%20Census%20of%20Ferry%20Operators&
DB_Short_Name=Ferry%20Census}
----> :link . [values]{http://www.data.gov/details/454}
```

[new]TRI.NET data engine for EPA Toxics Release Inventory

Figure 3. changes of datasets listed on data.gov

4. Conclusions

In the Data.gov Wiki, we have published billions of triples and built a couple of interesting applications/demos from gov-data. In most of our demos, a SPARQL web service played an important role in connecting distributed RDF data with conventional Web APIs. The simplicity of this process grants a promising future. Current efforts include converting more gov-data into RDF, developing more interesting applications and demonstrations that show how semantically linked government data can be used to combine information from the different datasets, looking at ways converted datasets can be linked with information found elsewhere on the Web, and examining linkages between potential US data.gov and UK linked gov-data efforts [4].

Acknowledgement: this work is partially supported by NSF Award #0524481, IARPA Award #FA8750-07-2-0031, DARPA award #FA8750-07-D-0185, #55-002001, #FA8650-06-C-7605.

5. REFERENCES

- [1] Berners-Lee, T., Linked Data, July 2007 <http://www.w3.org/DesignIssues/LinkedData.html>
- [2] Hendler, J. 2008. Web 3.0: Chicken Farms on the Semantic Web. *Computer* 41, 1 (Jan. 2008), 106-108
- [3] Berners-Lee, T., Putting Government Data online, June 2009 <http://www.w3.org/DesignIssues/GovData.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'07*, 2007