TWC LOGD:
A Portal for Linking Open Government Data

Dominic DiFranzo, Li Ding, John S. Erickson, Xian Li,
Tim Lebo, James Michaelis, Alvaro Graves, Gregory Todd Williams,
Jin Guang Zheng, Johanna Flores, Zhenning Shangguan, Gino Gervasio,
Deborah L. McGuinness and Jim Hendler

Tetherless World Constellation, Rensselaer Polytechnic Institute
{difrad, dingl, erickj4, lix15}@rpi.edu

Abstract. International open government initiatives are releasing an increasing number of "raw" government datasets directly to citizens via the Web, creating new opportunities but imposing burdens inherent to the problems of large-scale distributed data integration, collaborative data manipulation and transparent data consumption. With the goal of fostering a more scalable and interoperable Open Government Data ecosystem, the Tetherless World Constellation (TWC) team has developed the Linking Open Government Data (LOGD) Portal based on Semantic Web technologies. The TWC LOGD Portal is both an open source infrastructure supporting government data conversion, publishing, enhancement and access, and a vibrant community portal that educates and serves the growing international community of developers, data curators, and end users.

Keywords: Linked Data, Open Government Data, Semantic Web, Ecosystem

1 Introduction

In recent years the publication of Open Government Data (OGD) has grown more common around the world[1], emerging as a vital communications channel between governments and citizens. Persistent challenges faced by open government data stakeholders include the high cost of delivering large amounts of trusted data to the public as well as supporting the reuse and integration of data from disparate sources. Early solutions have focused on API-based architectures, resulting in isolated applications and visualizations that required expensive, specialized skills to develop. The Web of Data and open source semantic web technologies are well-suited for addressing the challenges of publishing and integrating Open Government Data [1, 2]. The TWC Linking Open Government Data (LOGD) Portal[2] now serves as a resource for the global community and makes this vision real.

1 http://www.data.gov/community/
2 http://logd.tw.rpi.edu/, referred as the TWC LOGD in the following sections.

2 TWC LOGD Portal Overview

The TWC LOGD Portal is a semantic web application dedicated to publishing Linked Data versions of OGD and sharing tools, services and expertise supporting an OGD ecosystem. It serves data users ranging from informed citizens, to domain experts, to developers consuming and creating novel applications enriched by government data.

2.1 Infrastructural Support for Open Government Data

The TWC LOGD Portal enables a Semantic Web-based ecosystem where users can actively convert, publish, access, archive, integrate and consume government-related data in connection with the Web of Data (see Figure 1). As of late September 2010 the Portal hosts more than 8.5 billion RDF triples from 436 RDFized datasets published by 11 different data sources, the majority of which are from data.gov.

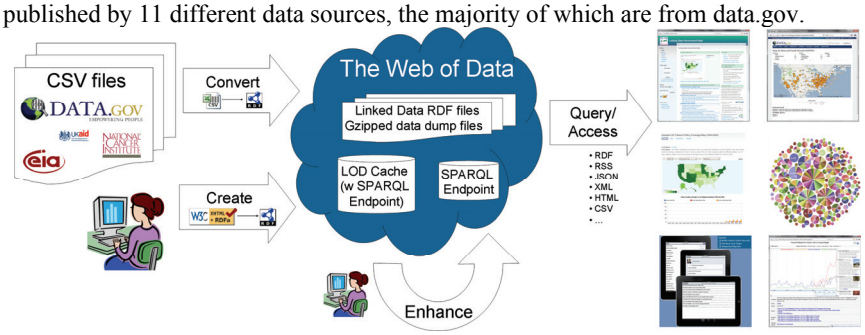


Figure 1. Data publishing workflow enabled by the TWC LOGD Portal

Data Conversion/Creation: Most government datasets are released in "raw" or unstructured formats. The TWC LOGD Portal converts these "raw datasets" to RDF using the TWC LOGD converter[3]. During conversion, the converter also captures metadata and provenance information critical for discovery, and it also maintains versions of datasets to ensure monotonic and persistent dataset publishing (i.e. new versions do not overwrite the old versions). The dataset versioning mechanism is used with a semantic diff (SemDiff) service [3] to compute changes in the Data.gov dataset catalog. [4] Semantic-capable content management systems, such as Semantic MediaWiki and Drupal with RDF modules, have been used to preserve user-generated metadata in normal content publishing activities using RDFa-annotated XHTML.

Data Enhancement: We also use social semantic web and machine learning technologies to enhance LOGD data with more links [3]. The TWC LOGD Portal has enhanced 54 converted datasets using 119 object properties, 65 classes and 7,051 links to other Linking Open Data (LOD) datasets[5]. Due in part to these enhancements, "TWC LOGD" appears in the September 2010 version of the LOD cloud diagram[6].

[3] http://data-gov.tw.rpi.edu/wiki/Csv2rdf4lod
[4] http://www.data.gov/raw/92/
[5] http://logd.tw.rpi.edu/twc-logd
[6] http://richard.cyganiak.de/2007/10/lod/imagemap.html

Data Query/Access: The converted datasets may be accessed through the TWC LOGD Portal in many ways. Each dataset has a summary web page that aggregates manually-contributed metadata (e.g. title, description, agency) and automatically-generated metadata (e.g. number of triples, links to data dumps). The metadata of datasets can also accessed by dereferencing URIs (Linked Data principle). In order to support users in accessing the datasets via query, there is also a publicly-accessible SPARQL endpoint[7] hosting a selection of datasets and the automatically recorded metadata of the TWC LOGD datasets. A TWC-developed tool - SparqlProxy[8] is used to enhance our SPARQL endpoint to return results in a richer set of formats such as JSON and HTML table. A TWC-developed tool - LOD cache[9] is used to synchronize RDF data published via RDFa-annotated web pages throughout the Portal.

2.2 Education and Community Portal

Each demo published through the TWC LOGD Portal is presented as an embedded, live mashup (or visualization) and a structured summary that describes the datasets, (semantic) technologies and SPARQL queries used. Comprehensive tutorials are provided to help developers quickly adopt the technologies demonstrated. RDFa on Drupal is used to publish both human- and machine-readable metadata embedded in TWC LOGD demos and tutorial pages; this approach hides the details of RDF creation from end users and enables a SPARQL-queryable site with novel features including the dynamically-generated list of demos on the TWC LOGD front page.

3 TWC LOGD Portal Highlights

Mashups are featured on the TWC LOGD Portal to demonstrate replicable coding practices for consuming government-related data in the Web of Data [4]. Figure 2 illustrates a four-step workflow of a TWC LOGD demo - "CASTNET Ozone Map"[10]. Steps 2-4 highlight three levels of mashup: *data*, in which applications combine datasets from different sources (or triple stores) using SPARQL queries; *visualization*, in which applications leverage multiple visualization libraries and APIs such as the Google Visualization API and MIT SIMILE Exhibit; and *application*, in which related applications created by different parties are interlinked using HTTP URLs.

3.1 Mashups and Community Impact

The TWC LOGD Portal demonstrates how Linked Data principles and semantic web technologies may be applied to decrease development costs and increase the reuse of

[7] http://logd.tw.rpi.edu/sparql
[8] http://logd.tw.rpi.edu/tutorial/how_to_use_sparqlproxy
[9] http://data-gov.tw.rpi.edu/ws/lodcx.php
[10] http://logd.tw.rpi.edu/demo/clean_air_status_and_trends_-_ozone

data models, links and visualization techniques. It advocates a bottom-up approach, encouraging developers to collaboratively model data, define terms, link terms and concepts to other heterogeneous datasets, and to use generic visualization libraries and APIs to more quickly get useful applications running. For example, the "CASTNET Ozone Map" (see Figure 2) , which mashes up multiple LOGD datasets and Web-based visualization APIs, was created within two weeks and iteratively enhanced over the span of a month. Similarly, in September 2010 four unique demos were created to support Tobacco Prevalence study in the NIH project PopSciGrid[11].

Developers don't need to be expert in semantic technologies or Linked Data principles to create semantically-enabled LOGD mashups. Undergraduate students in RPI's Fall 2009 Web Science class created mashups using semantic technologies and datasets found on the TWC LOGD Portal. Given a two-hour introduction to basics like RDF and SPARQL and patterns for using visualization tools like the Google Visualization API, each group created visualizations mashing at least two converted datasets in less than two weeks. Similarly, in August 2010 the US Data.gov project hosted a Mash-a-thon workshop[12], organized in part by TWC to engage government developers and data curators in hands-on learning using TWC LOGD tools and datasets. In just two days four teams successfully built LOGD-based mashups, demonstrating the low cost of knowledge transfer and the rapid learning process inherent in the best practices embodied by the TWC LOGD Portal.

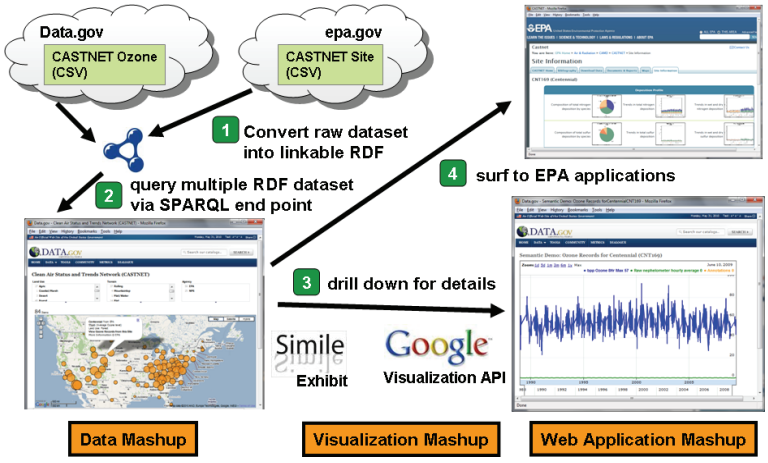


Figure 2. Data Mashup workflows as exemplified by the CASTNET Ozone Map

3.2 Mashups and Innovative Technologies

Data integration is vital to the objectives of OGD. Linking and integrating data in novel ways help consumers uncover new patterns and correlations and create new

[11] http://apha.confex.com/apha/138am/webprogram/Paper223359.html
[12] http://logd.tw.rpi.edu/mashathon2010/

knowledge. Linked Data principles and semantic web technologies make it easy to connect heterogeneous datasets without advance coordination or planning.

The TWC LOGD Portal landing page is itself a mashup of data from multiple sources. As shown in Figure 3, the content panels on the TWC LOGD front page are based on live SPARQL queries across the site data using XSLT and the Google Ajax API. The sources include metadata of the TWC LOGD datasets stored at the TWC LOGD Data triple store; metadata of demos and tutorials published via RDFa-annotated XHTML pages that are synchronized with the LOD Cache; relevant news items maintained in the Data-Gov website[13]; and RSS feeds from the TWC LOGD Portal, Data-Gov's SemDiff Service and the Google News site.

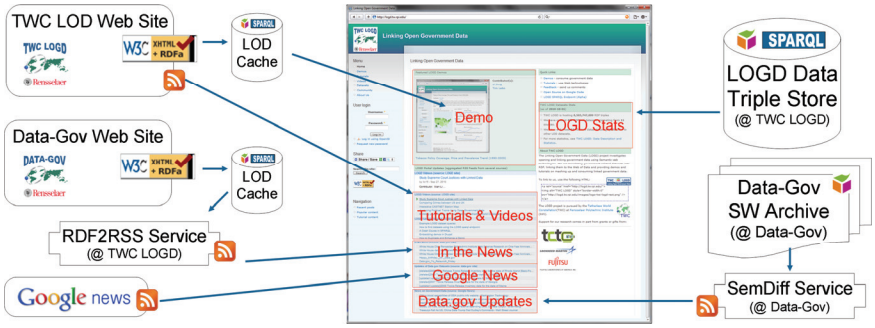


Figure 3. The TWC LOGD Portal landing page as a dynamically-sourced mashup

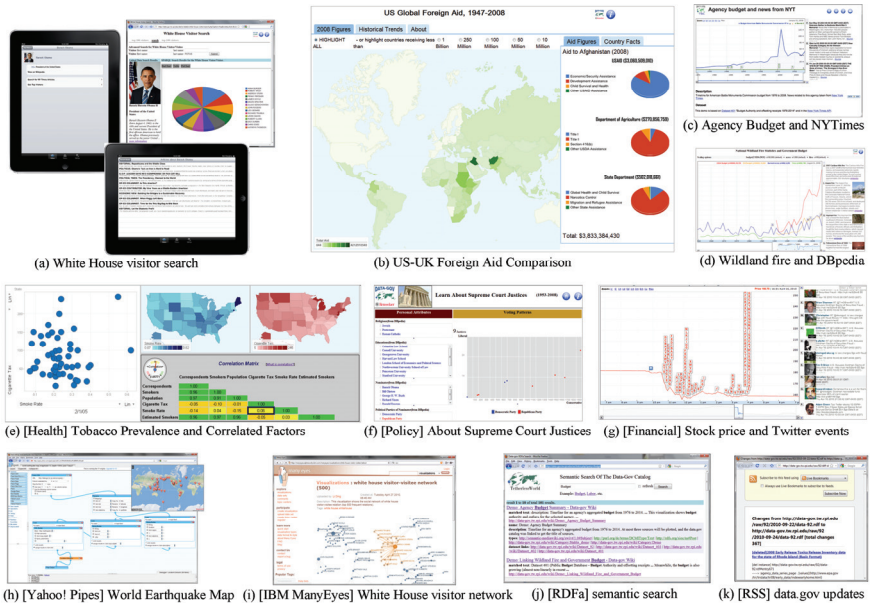
A team graduate and undergraduate students have created over 40 different mashups and visualizations on the TWC LOGD Portal. These mashups are diverse, including demonstrating the integration of data from multiple sources including DBpedia, the New York Times API, and open government data produced by non-US sources (see Figure 4 a-g); deploying data via web and mobile interfaces (see Figure 4 a); supporting interactive analysis for specific domains including health, policy and financial data (see Figure 4 e-g); consuming integrated data using readily-available Web-based services (see Figure 4 h,i); and designing data access tools (Figure 4 j) and semantic data integration tools (Figure 4 k).

3.3 Mashups, Transparency and Provenance

As data is processed, converted, enhanced, and republished consumers are further removed from the original once-raw form. For data published through the TWC LOGD Portal this is especially true because it is curated by a university aggregating data from a variety of government and non-government sources, each with its own degree of authority and trustworthiness. The TWC LOGD data conversion processes capture the lineage of the data products published, enabling data consumers to inspect and query dataset metadata and provenance and thus perceive enhanced credibility for derived data products and applications.

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

The TWC LOGD demo "Provenance Mashup"[14] summarizes when Data.gov datasets were last updated as determined by their HTTP header responses. This is a step toward revealing the temporal provenance metadata of the datasets that is not clearly stated in the Data.gov dataset catalog. Since the TWC LOGD converter captures this along with other provenance information and serializes it in Proof Markup Language (PML) [5], consumers are able to query this metadata and display and utilize it in applications and visualizations.



(a) White House visitor search (b) US-UK Foreign Aid Comparison (c) Agency Budget and NYTimes (d) Wildland fire and DBpedia

(e) [Health] Tobacco Prevalence and Correlated Factors (f) [Policy] About Supreme Court Justices (g) [Financial] Stock price and Twitter events

(h) [Yahoo! Pipes] World Earthquake Map (i) [IBM ManyEyes] White House visitor network (j) [RDFa] semantic search (k) [RSS] data.gov updates

Figure 4. A selection of TWC LOGD mashups and visualizations

4 Conclusions and Future Work

The TWC LOGD Portal represents a significant advancement for global OGD objectives but there are many ways it can be extended to more fully serve the LOGD community. For example, the Portal should interactively engage users through datasets, demos and tutorial-centered discussion threads, applying semantic web technologies to integrate relevant topics across the portal. The recent TWC-led government Mash-a-thon highlighted the value of interaction between the participants with the TWC LOGD team; we perceive the Portal as a 24x7x365, community-driven extension of that interaction model.

The publication of converted government datasets is a critical service of the TWC LOGD Portal, thus we plan to significantly extend the scope of our Data.gov dataset conversions. In particular, TWC will soon add 270K US government geodata datasets

[14] http://logd.tw.rpi.edu/demo/datagov_survey_dataset_modification_dates

to the TWC LOGD conversion workflow, accompanied by relevant demos and tutorials to facilitate consumption and reuse of this new class of data. Further, we are working on greater encoding and exposing of provenance information in addition to creating tools for allowing user-contributed annotation of demos so users can help identify potential data questions or updates.

The LOGD world is vast, represented by diverse stakeholders ranging from providers, curators, and developers, to civil servants, activists, community leaders, average citizens and beyond; the long-term goal is for the TWC LOGD Portal to become the focal point for engaged discussion and outreach centered on LOGD issues, technologies and best practices.

References

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

[2] 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)

[3] 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 AAAI Spring Symposium on Linked Data Meets Artificial Intelligence. (2010)

[4] Ding,L., McGuinness,D., Hendler,J., Mash-up of Linked Government Data from http://data.gov, In Proceedings of SemTech 2010, (2010)

[5] McGuinness, D. and Pinheiro da Silva, P. Explaining Answers from the Semantic Web: The Inference Web Approach. Journal of Web Semantics. 1(4). (2004)

Appendix: SWC2010 Report

1. ***Does the TWC LOGD Portal meet the SWC2010 "Minimum Requirements?"***

1.1 *"end user application"* Yes. 40+ demos deliver LOGD and technologies to citizens (e.g. White House visitor search, US/UK foreign aid), domain experts (e.g. PopSciGrid demos) and open source developers (e.g. RSS of data.gov updates). We have 400K page visits from 134 countries and 4K cities. See **Section 2**

1.2.a. *"diverse ownership or control"* Yes, owners/controllers include numerous government agencies and other social entities (e.g. DBpedia, New York Times, Twitter, Google Search). See **Sections 2** and **3.2**

1.2.b. *"heterogeneous sources"* Yes, e.g. our demos and TWC LOGD landing page consume multiple triple stores, semantic data files and Web Data APIs, Our converter also RDFized raw data in CSV, XML and Fixed-width text. See **Section 3**

1.2.c. *"real world data"* Yes; see **Section 2.** Therefore, TWC LOGD has been included in the Sept 2010 version of the *Linked Open Data Cloud*.

1.3.a. *"using Semantic Web technologies"* Yes, e.g. owl:sameAs used for linking datasets by state, ontology for dataset/demo metadata; semantic diff used for computing changes of RDF data. See **Sections 2.1** and **3.2**

1.3.b. *"data manipulated/processed in interesting ways"* Yes, e.g. we leverage RDFa to preserve structure metadata in a content management system and then support a queryable Web using a LOD cache. We also provide powerful data enhancement functions, e.g. promoting literal strings into DBpedia URI, and enable cell-based conversion for multi-dimension data tables. Our mashups help multiple dataset analysis across domains and times so as to reveal hidden facts (or stimulate hypotheses) which are impossible within a single dataset. See **Section 3.2**.

1.3.c. *"central role in achieving things that alternative technologies cannot do as well.* Yes; RDF and SPARQL are critical to integration. See **Sections 2.1** and **3.2**.

2. ***What "Additional Desirable Features" does TWC LOGD demonstrate?***

The TWC LOGD Portal extends Drupal 6 with semantic technologies to create a flexible, scalable collaboration environment with ***an attractive and functional Web interface***. Portal content is ***dynamically presented*** using a combination of Drupal, XSLT, SPARQL and RDFa. SPARQL is also used to query external data and present results to users. Most Portal pages have "Like" (enabled by the Open Graph Protocol and RDFa) and "Rate" buttons, enabling end users to give feedback.

The ***scalability*** of the TWC LOGD infrastructure is demonstrated by the large and diverse datasets being actively converted and published on a daily basis. TWC LOGD mashups also exhibit how distributed services and data sources may be integrated using semantic technologies. Our open source code presents best practices for combining dynamic and static data in scalable real-time visualizations.

In addition to interactive demos of data mashups, the TWC LOGD Portal hosts ***multimedia documents*** including videos and slides to help stakeholders understand demos and tutorials.

The unique work represented by the TWC LOGD Portal has been ***rigorously evaluated*** in the US Government and discussed in the White House blog, commending its use of Data.gov datasets in innovative ways to generate practical applications and mashups. Details of TWC's role and the impact of applying semantic web technologies can be found in **Sections 3.1** and **3.2**.

TWC is actively engaged with organizations inside and outside of government, and this enables us to receive and act on feedback concerning our data conversion, tools, and applications. **Section 3.3** details this interaction. Diagrams in **Sections 2.1** and **2.2** illustrate the TWC LOGD Portal's ***data workflow***. **Section 3.3** shows additional work on data and provenance that ***goes beyond pure information retrieval***.

The TWC LOGD Portal uses several approaches to ensure the ***accuracy of results*** by improving the quality of converted data. Evolving heuristics based on statistical analysis are used to connect entities with the same meaning across massive data, such as linking datasets based on US States.

The TWC LOGD Portal currently acts like a research-oriented, US-based site. It provides ***diversified accessibility*** including the "White House Visitor" demo on iPad, submitted into iTunes store for distribution. No alternate language versions of our content or links to translation services are currently available.