TWC LOGD: A Portal for Linked Open Government Data Ecosystems☆

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

Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180, USA

Abstract

International open government initiatives are releasing an increasing volume of raw government datasets directly to citizens via the Web. The transparency resulting from these releases creates new application opportunities but also imposes new burdens inherent to large-scale distributed data integration, collaborative data manipulation and transparent data consumption. The Tetherless World Constellation (TWC) at Rensselaer Polytechnic Institute (RPI) has developed the Semantic Web-based TWC LOGD Portal to support the deployment of Linked Open Government Data (LOGD). The Portal is both an open source infrastructure supporting linked open government data production and consumption and a vibrant community portal that educates and serves the growing international open government community of developers, data curators and end users. This paper motivates and introduces the TWC LOGD portal and highlights innovative aspects and lessons learned.

Keywords: Linked Data, Open Government Data, Ecosystem, Data.gov

1. Introduction

In recent years we have observed a steady growth of Open Government Data (OGD) publication, emerging as a vital communication channel between governments and their citizens. A number of national and international Web portals (e.g., Data.gov and Data.gov.uk[1]) have been deployed to release OGD datasets online. These datasets embody a wide range of information significant to our daily lives, e.g., locations of toxic waste dumps, regional health-care costs and local government spending. A study conducted by the Pew Internet and American Life Project reported that 40% of adults went online in 2009 to access government data [1]. One direct benefit of OGD is richer governmental transparency: citizens are now able to access the raw government data behind the previously-opaque applications. Rather than being merely "read-only" users, citizens can now participate in collaborative government data access, including "mashing up" distributed government data from different agencies, discovering interesting patterns, customizing applications, and providing feedback to enhance the quality of published government data.

For governments, the costs of providing data are reduced when the data is released through these OGD portals as opposed to rendered into reports or applications. However, for users of the data, this can cause interoperability, scalability and usability problems. OGD raw datasets are typically available *as is* (i.e., in heterogeneous structures and formats), requiring substantial human workload to clean them up for machine processing and to make them comprehensible. To accelerate the usage of government data by citizens and developers, we need an effective infrastructure with more computing

☆The work in this paper was supported by grants from the National Science Foundation, DARPA, National Institute of Health, Microsoft Research Laboratories, Lockheed Martin Advanced Technology Laboratories, Fujitsu Laboratories of America and LGS Bell Labs Innovations. Details of the support can be found on the TWC LOGD Portal.

*Corresponding author at: Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180, USA

[1]An ongoing list of countries with OGD portals is provided via `http://www.data.gov/opendatasites`

power to process big OGD data and better social mechanisms to distribute the necessary human workload to stakeholder communities.

Recent approaches, such as Socrata[2] and Microsoft's OData[3], advocate distributed RESTful data APIs. These APIs, however, only offer restricted access to the underlining data through their pre-defined interfaces and can introduce non-trivial service maintenance costs. The emerging Linked Open Government Data (LOGD) approach [2, 3, 4], which is based on Linked Data [5] and Semantic Web technologies, overcomes these limitations on data reuse and integration. Instead of providing data access APIs based on some assumed requirements, the LOGD approach directly exposes OGD datasets to consumers in Linked Data representation via e.g., RDF dump files and SPARQL endpoints. The open nature of LOGD supports incrementally interlinking OGD datasets with other datasets. Moreover, the Web presence of LOGD allows developers to access data integration results (e.g., SPARQL query results) in JSON and XML, making it easy to build online data mashup applications which are good incentives for LOGD adoption.

The LOGD approach has recently been promoted by a combination of government and academic thought leaders in both the US and the UK. In particular, LOGD has been deployed at Data.gov.uk in a top-down style, i.e., mandating OGD datasets to be published in RDF, while in the US LOGD has been deployed through Data.gov in a bottom-up style, i.e., RPI's TWC LOGD project has converted Data.gov datasets into RDF and the knowledge was then transferred to Data.gov. This paper describes how the TWC LOGD Portal[4] has been designed and deployed from the ground-up to serve as a resource for the global LOGD community and make LOGD deployed in the US. This work contributes at multiple levels: it demonstrates practical applications of Linked Data in publishing and consuming OGD data; it represents the first Semantic Web platform to play a role in US open government activities (`http://data.gov/semantic`), and, as we will discuss later in this paper, it contributed the largest meaningful real world dataset in the Linking Open Data (LOD) cloud[5] to date.

In the remainder of this paper, we provide an overview of the TWC LOGD Portal, review our system design for LOGD production and consumption, discuss provenance and scalability issues in the LOGD community, and conclude with future directions.

[2] `http://opendata.socrata.com`
[3] `http://www.odata.org`
[4] `http://logd.tw.rpi.edu`
[5] `http://richard.cyganiak.de/2007/10/lod`

2. Overview of the TWC LOGD Portal

We define a *LOGD ecosystem* as a Linked Data-based system where stakeholders of different sizes and roles find, manage, archive, publish, reuse, integrate, mashup, and consume open government data in connection with online tools, services and societies. An effective LOGD ecosystem serves a wide range of users including government employees who curate raw government data, developers who build applications consuming government data, and informed citizens who view visualizations and analytical results from government data. The TWC LOGD Portal provides a key infrastructure in support of LOGD ecosystems. Figure 1 shows the high-level workflow embodied by the Portal to meet the critical challenges of supporting large-scale LOGD production, promoting LOGD consumption and growing the LOGD community.

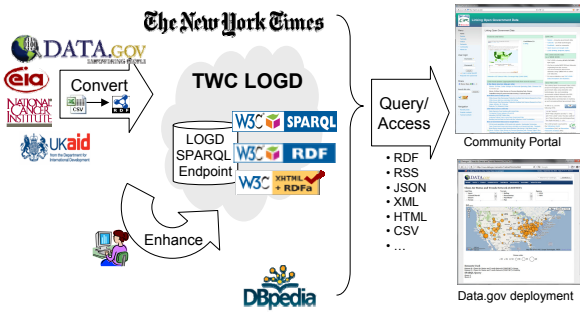

Figure 1: The High-level Workflow of the TWC LOGD Portal

LOGD Production: Grounding LOGD deployment on a critical mass of real world OGD datasets requires an effective data management infrastructure. We have therefore developed a data organization model with tools to enable a fast, persistent and extensible LOGD production infrastructure. The LOGD data produced by this infrastructure has been adopted by Data.gov and was linked into the global LOD cloud in 2010.

LOGD Consumption: The adoption of LOGD depends on its perceived value as evidenced by compelling LOGD-based end user applications. Over 50 live online demos have been built and hosted on the Portal, using a wide range of web technologies including data visualization APIs and web service composition.

LOGD Community: The growth of LOGD ecosystems demands active community participation. We have therefore added collaboration and education mechanisms to the Portal to support knowledge sharing and promote best practices in the LOGD community. We have also enriched transparency by declaratively tracing the provenance of LOGD workflows.

3. LOGD Production

Published OGD datasets often have issues that impede machine consumption, e.g., proprietary formats, ambiguous string-based entity reference and incomplete metadata. This section shows how the TWC LOGD Portal addresses these difficulties in LOGD production[6].

3.1. LOGD Data Organization Model and Metadata

In order to support users to access data at different levels of granularity and to maintain persistent data access, we defined a data organization model built around the publishing stages and the structural granularity of LOGD datasets. This model is used to design Linked Data URIs. In what follows, we use Data.gov *Dataset 1623*[7] to exemplify the model.

3.1.1. Data Publishing Stages

Focusing on persistency, we identify three *data publishing stages* to support unfettered growth of LOGD, i.e., (i) any dataset additions or revisions will be incrementally added without changing existing data, and (ii) every dataset, dataset version, and dataset conversion result has its own permanent URI.

At the **catalog stage**, we create an inventory of *datasets*, i.e., online OGD datasets, for LOGD production. In the US, each Data.gov dataset is published by a certain government agency with a unique numerical identifier and the corresponding metadata. For example, *Dataset 1623* is released by the US Department of Health and Human Services and contains information about Medicare claims in US states. The identity of a dataset contains two parts: the *source_id* that uniquely identifies the source of the dataset[8] and the *dataset_id* that uniquely identifies the dataset within its source. In our example, we use `http://logd.tw.rpi.edu` as *base_uri*, "data-gov" as *source_id*, and "1623" as *dataset_id*. Dereferencing a URI in the example below will return either a web page with RDFa annotation or an RDF/XML document depending on HTTP content negotiation. The metadata of a dataset shows the type, identifier, metadata web page and modification date of the dataset, and it also includes links to the source and subsets of the dataset. While many datasets are provided as a single file, others contain multiple files. For example, *Dataset 1033*[9] uses separate files to describe people, facilities and organizations. Therefore, we include an extra part[10] in the dataset's identifier and a new level in the corresponding *void:subset* hierarchy so that we can distinguish data associated with different files.

```
Syntax:
  <source_uri>        ::= <base_uri> "/source/" <source_identifier>
  <dataset_whole_uri> ::= <source_uri> "/dataset/" <dataset_identifier>
  <dataset_part_uri>  ::= <dataset_whole_uri> "/" <part_identifier>
  <dataset_uri>       ::= <dataset_whole_uri> | <dataset_part_uri>

Example URIs:
  http://logd.tw.rpi.edu/source/data-gov
  http://logd.tw.rpi.edu/source/data-gov/dataset/1033
  http://logd.tw.rpi.edu/source/data-gov/dataset/1033/fm_facility_file
  http://logd.tw.rpi.edu/source/data-gov/dataset/1623

Example Metadata (Dataset 1623):
@prefix conversion: <http://purl.org/twc/vocab/conversion/> .
@prefix void:       <http://rdfs.org/ns/void#> .
@prefix foaf:       <http://xmlns.com/foaf/0.1/> .
@prefix dcterms:    <http://purl.org/dc/terms/> .
@prefix xsd:        <http://www.w3.org/2001/XMLSchema/> .

<http://logd.tw.rpi.edu/source/data-gov/dataset/1623>
  a void:Dataset , conversion:AbstractDataset ;
  conversion:base_uri           "http://logd.tw.rpi.edu" ;
  conversion:source_identifier  "data-gov" ;
  conversion:dataset_identifier "1623" ;
  dcterms:identifier            "data-gov 1623" ;
  dcterms:contributor <http://logd.tw.rpi.edu/source/data-gov> ;
  foaf:isPrimaryTopicOf
    <http://logd.tw.rpi.edu/source/data-gov/dataset_page/1623> ;
  void:subset
    <http://logd.tw.rpi.edu/source/data-gov/dataset/1623/version/2010-Sept-17>,
    <http://logd.tw.rpi.edu/source/data-gov/dataset/1623/subset/meta> ;
  dcterms:modified "2010-09-09T12:32:49.632-00:05"^^xsd:dateTime .
```

At the **retrieval stage**, we create a *dataset version*, i.e., a snapshot of the dataset's online data file(s) downloaded at a certain time, and use it as the input to our LOGD converter. The URI of a dataset version depends on the URI of the corresponding dataset. The metadata of a dataset version links to the corresponding dataset, subsequent conversion layers, and a dump file containing RDF triples converted from the version.

```
Syntax:
  <version_uri> ::= <dataset_uri> "/version/" <version_identifier>

Example URI:
  http://logd.tw.rpi.edu/source/data-gov/dataset/1623/version/2010-Sept-17
```

At the **conversion stage**, we create configurations and convert a dataset version to *conversion layers*, each of which is a LOGD representation of the version. The basic conversion configuration, called "raw", is automatically created by the Portal. It minimizes the need for user input when converting data tables to RDF and preserves table cell content as strings [6]. Users can add more enhancement configurations to increase the quality of LOGD, e.g., promoting named entities to URIs

[6] In this paper we focus on US datasets that are described in English. We are currently working on multilingual support for international datasets - see `http://logd.tw.rpi.edu/demo/international_dataset_catalog_search`

[7] `http://www.data.gov/details/1623`, OMH Claims Listed by State

[8] A source could be a person or an organization. Although an arbitrary string can be used to identify a source organization, we recommend using the host name of its website, e.g., use "epa-gov" for EPA `http://epa.gov`.

[9] `http://www.data.gov/details/1033`, EPA FRS Facilities Combined File CSV Download for the Federated States of Micronesia

[10] conversion:subject_discriminator is used to provide this identifier. We recommend using the file name to name the dataset part.

and mapping *ad hoc* column names to common properties [7]. A conversion layer has a *conversion identifier* in the form of "enhancement/N", where *N* is an integer. Each conversion layer is generated using a unique configuration, reflecting an independent semantic interpretation of the version, and physically stored in its own dump file. The conversion layers of a dataset version can be interlinked by describing the same table rows and enhancing the same table columns. The URI of a conversion layer depends on the corresponding dataset, dataset version and configuration. Its metadata connects the conversion layer to e.g., the corresponding dataset version and a dump file containing RDF triples generated by the conversion. Simple statistics of the layer are also provided, including a list of properties, a list of sample entity URIs and the number of triples generated.

```
Syntax:
  <conversion_uri> ::= <version_uri> "/conversion/" <conversion_identifier>

Example URIs:
 http://logd.tw.rpi.edu/source/data-gov/dataset/1623
         /version/2010-Sept-17/conversion/raw

 http://logd.tw.rpi.edu/source/data-gov/dataset/1623
         /version/2010-Sept-17/conversion/enhancement/1
```

3.1.2. Data Structural Granularity

We also allow consumers to link and access LOGD datasets at different levels of structural granularity.

Data Table: Data tables (e.g., relational database and Excel Spreadsheet) are widely used by government agencies in publishing OGD raw datasets[11]. A data table is identified by the corresponding version URI.

Record and Property: A data table contains rows and columns, each column representing a particular *property*, each row corresponding to a *record*, and each table cell storing the actual value of the corresponding property in the corresponding record. In the Portal, properties and records are identified by automatically generated URIs[12] and the value in a table cell is usually represented by an RDF triple in the form of (record_uri, property_uri, cell_value)[13]. The URI of a property is independent from *version_id* because we assume that the meaning of a column, which maps to a property, will remain the same in all versions of a dataset. The URI of a record is independent from *conversion_id* to facilitate mashing up descriptions of the same record from different conversion layers of a version.

[11] We leave non-tabular structures, e.g., XML trees, to future work.

[12] The name of a property is derived from the header name of the corresponding column: turning non-alpha-numerical character sequences into one underscore character and trimming the heading and tailing underscore characters of the result. A row number is a positive number that starts from 1.

[13] Advanced conversion may even assign a URI to a cell.

```
Syntax:
  <property_uri> ::= <dataset_uri> "/vocab/" <conversion_id> "/" <property_name>
  <record_uri>   ::= <version_uri> "/thing_" <row_number>

Example URIs (property URI and record URI):
  http://logd.tw.rpi.edu/source/data-gov/dataset/1623/vocab/raw/state
  http://logd.tw.rpi.edu/source/data-gov/dataset/1623/version/2010-Sept-17/thing_32
```

Entity and Class: A record can mention named entities such as people, organizations and locations. Our LOGD converter supports the promotion of string-based identifiers to URIs and the creation of *owl:sameAs* mappings to other URIs. The corresponding property is promoted to an *owl:ObjectProperty*, and the entity may also be typed to an automatically-generated class[14]. The automatically generated properties and classes are local to the dataset. This allows third parties to create heuristic algorithms suggesting ontology mappings across different datasets. Users can therefore query multiple LOGD datasets which share mapped properties and classes.

The following example shows the URI for the state "Arkansas" within dataset 1623 and the corresponding metadata generated in enhancement conversion. An *owl:sameAs* statement links the local URI to the corresponding DBpedia URI, making dataset 1623 part of the LOD cloud.

```
Syntax:
  <entity_uri> ::= <dataset_uri> "/typed/" <class_name> "/" <entity_name>
  <class_uri>  ::= <dataset_uri> "/vocab/" <class_name>

Example URIs (entity and class respectively):
  http://logd.tw.rpi.edu/source/data-gov/dataset/1623/typed/state/Arkansas
  http://logd.tw.rpi.edu/source/data-gov/dataset/1623/vocab/State

Example Metadata (the entity of "Arkansas" in Dataset 1623):
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl:  <http://www.w3.org/2002/07/owl#> .
<http://logd.tw.rpi.edu/source/data-gov/dataset/1623/typed/state/Arkansas>
  a  <http://logd.tw.rpi.edu/source/data-gov/dataset/1623/vocab/State>;
  rdfs:label  "Arkansas" ;
  owl:sameAs <http://dbpedia.org/resource/Arkansas>,
             <http://sws.geonames.org/4099753> .
```

3.2. LOGD Production Workflow

The LOGD production workflow forms the centerpiece of the TWC LOGD Portal and decomposes into a sequence of critical steps. These steps balance human intervention and automation in order to achieve a functionally correct yet efficient workflow:

- Initialize: given an online OGD dataset, determine its source_id, dataset_id, title, homepage, and the URIs from which the raw data can be downloaded.
- Retrieve: create a *version_id* upon generating a dataset version. Construct a file folder on the TWC LOGD Portal and download the corresponding raw data (and optionally the auxiliary documents) to build an archival snapshot.

[14] The local name of the class URI is provided as an enhancement parameter.

- Cleanup: cleanup the archived version of the OGD dataset into a collection of CSV files[15], each containing one data table.
- Convert: convert tabular data (in CSV files) into LOGD (in RDF files) using the automatically generated raw conversion configuration.
- Enhance: optionally, use user-contributed enhancement conversion configurations to generate additional conversion layers.
- Publish: publish LOGD datasets on the Web as (i) downloadable RDF dump files and (ii) dereferenceable Linked Data (backed by the TWC LOGD SPARQL endpoint).

The workflow is currently operated by the TWC LOGD team using our open source tool named `csv2rdf4lod`[16]. Most workflow steps (e.g., retrieve, convert and publish) have been mapped to automated scripts on the server side.

3.3. Links in LOGD

Linking data at different levels of structural granularity. As discussed in section 3.1, we create unique URIs for LOGD data at different levels of granularity to support linking. A property from the enhancement conversion layer links to the corresponding property in the raw conversion layer via *conversion:enhances*. A record links to the corresponding dataset version via *dcterms:isReferencedBy* and *void:inDataset*, and to its referenced entities via dataset-specific properties. A conversion layer links to the properties it uses via *conversion:uses_predicate*, and to its sample records via *void:exampleResource*.

Linking data at different publishing stages. The data identified at different publishing stages are all considered as instances of *void:Dataset*. We further connect them using *void:subset* relations, e.g., a version is a *void:subset* of a dataset.

Linking data via provenance traces. We also use the Proof Markup Language (PML) [8] to link datasets based on data retrieval and data derivation operations. For example, we use a "pcurl" script from `csv2rdf4lod` to download raw data from the Web and create a version. This script also generates PML metadata linking the local files to the original web URIs. Similarly, we capture provenance trace when converting from local CSV files into LOGD RDF files and when loading those files into our SPARQL endpoint.

Linking data via owl:sameAs. We have investigated statistical methods as well as social semantic web based methods [4] to establish *owl:sameAs* links from entities recognized in LOGD data to other LOD datasets. For instance, we use heuristic algorithms to identify the occurrences of US states in LOGD datasets, and then map them to the corresponding DBpedia URIs [9].

Linking data via shared/linked ontological terms. The Data-gov Wiki[17], which is based on Semantic MediaWiki [10], provided a social semantic web platform for users to link dataset specific terms to more general terms [4]. For example, the URI `http://data-gov.tw.rpi.edu/vocab/p/92/title` will be dereferenced to a RDF/XML document exported from a wiki page, where users can add more semantic definition, e.g., this property is a subproperty of *rdfs:label*. This link allows Linked Data browsers such as Tabulator [11] to correctly interpret the property. With the help of `csv2rdf4lod`, the current TWC LOGD Portal allows users to directly add common properties/classes using conversion configurations [7].

3.4. Achievements and Community Impact

The TWC LOGD Portal maintains metadata for every dataset, dataset version and conversion layer. Starting from a dataset catalog page[18], the metadata can be accessed on the Portal in the form of both web pages and Linked Data. All metadata and some LOGD datasets have been loaded into the TWC LOGD SPARQL endpoint[19], which is powered by OpenLink Virtuoso[20].

To date, our work has contributed to several different communities of practice. *US Government*: Semantic Web technologies have been deployed in the US government for open government data: 6.4 billion RDF triples of our converted data have been published at `http://www.data.gov/semantic`[21], RDF files are available for download for some government datasets listed at Data.gov, and a SPARQL endpoint (based on a Virtuoso triple store) serving some of these triples is available at `http://services.data.gov/sparql`. *Linked Data Community*: As of May 2011 the TWC LOGD Portal hosts more than 9.9 billion RDF triples

[15] At the current stage, the TWC LOGD Portal focuses on tabular government data. OGD data in other formats are converted into CSV.

[16] `http://logd.tw.rpi.edu/technology/csv2rdf4lod`

[17] `http://data-gov.tw.rpi.edu`, it is the previous implementation of the TWC LOGD Portal.

[18] `http://logd.tw.rpi.edu/datasets`

[19] `http://logd.tw.rpi.edu/sparql`

[20] `http://virtuoso.openlinksw.com`

[21] This older data was generated using a previous converter [6].

from 1,838 OGD datasets published by 82 different data sources, and most datasets are from Data.gov[22]. The scale and governmental composition of this collection have made "TWC LOGD" (`http://logd.tw.rpi.edu/twc-logd`) the largest meaningful *real world* dataset in the LOD cloud to date. The Portal has enhanced 1,505 datasets, and we have accumulated 8,335 *owl:sameAs* statements for 37 datasets (including 25 Data.gov datasets) linking to LOD datasets such as DBpedia, GeoNames and GovTrack. *Open Source community*: We have released `csv2rdf4lod` as open source project on GitHub. We are currently working with several community-based organizations to teach them how to create and exploit LOGD directly from their local government data assets.

4. LOGD Mashups

Mashups featured on the TWC LOGD Portal demonstrate replicable coding practices for consuming LOGD datasets. Although individual government datasets may contain useful content, applications based on open government data are much more interesting and useful when they mash up datasets from a variety of sources, especially from inside and outside the government. For example, one demonstration explores correlations between Medicare claim numbers and state adjusted gross income[23], while another uses national wildfire statistics and Wikipedia's famous wildfires information to evaluate the government budget on wildfire fighting[24].

4.1. LOGD Mashup Workflow

In order to understand the workflow for building a LOGD mashup, we discuss the key steps using the White House Visitor Search[25] application as an example. Figure 2 illustrates how the application mashes up information about employees of the White House (e.g., the President of the United States) from multiple sources including descriptions and photos from *DBpedia* (which extracts Wikipedia Infobox content into Linked Data), visitor statistics (displayed in a pie chart) from a LOGD dataset *White House Visitor Records*, which was converted from the White House Visitor Record[26], and the name mappings (between the above two datasets) from another LOGD dataset *White House Visitee to DBpedia link*[27], which maintains user-contributed knowledge on the Data-gov Wiki.

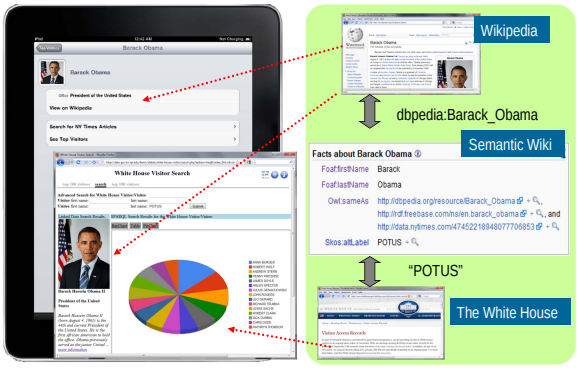

Figure 2: An Example LOGD Mashup: White House Visitor Search

- Load Data: Before building a mashup, users often need to load the distributed datasets into SPARQL endpoints for efficient data integration[28]. In our example, the three datasets above were loaded into two different SPARQL endpoints: *DBpedia* dataset was loaded into the DBpedia SPARQL endpoint[29], the other two datasets were loaded into the TWC LOGD SPARQL endpoint.

- Request Data: An application composes SPARQL queries on several datasets and issues the queries to related SPARQL endpoint(s). In our example, SPARQL queries are dynamically constructed for each individual White House Visitee. One SPARQL query is first issued to the TWC LOGD SPARQL endpoint to integrate the *White House Visitor Record* dataset and the *White House visitee to DBpedia link* dataset (joined by the same first name and last name), and then, a DBpedia URI retrieved from the query, if available, is used to compose another SPARQL query to request the person's biographical information from the DBpedia SPARQL endpoint.

- Convert the Results: The SPARQL query results (in SPARQL/XML) are then converted to various formats so that they can be consumed as web data

[22] This newer data was generated using `csv2rdf4lod`.
[23] `http://logd.tw.rpi.edu/node/21`
[24] `http://logd.tw.rpi.edu/node/57`
[25] `http://logd.tw.rpi.edu/demo/white-house-visit/search`
[26] `http://www.whitehouse.gov/briefing-room/disclosures/visitor-records`
[27] `http://data-gov.tw.rpi.edu/wiki/White_House_Visitee`
[28] Users may direct dereference Linked Data on the fly, but preloading SPARQL endpoint meets reliability and efficiency requirements from applications.
[29] `http://dbpedia.org/sparql`

APIs. In our example, SPARQL query results are consumed in SPARQL/JSON format[30], converted directly by our Virtuoso-based SPARQL endpoint (we also provide an open source tool called Sparql-Proxy[31] to convert SPARQL query results to other frequently used formats).

- Integrate the Results: Upon receiving multiple results from SPARQL endpoint(s), a LOGD application uses shared URIs (or other unique string identifiers) to link the query results. In our example, the SPARQL query results from the two SPARQL endpoints are integrated by common DBpedia URIs.

- Visualize the Results: The LOGD application finally presents the data mashup to the end users via maps, charts, timelines and other visualizations using Web-based visualization APIs. In our example, the Google Visualization API[32] and jQuery[33] are used to enable an AJAX-style UI.

4.2. Mashups and Innovative Technologies

Linking and integrating data is critical for consumers to uncover new correlations and create new knowledge in government data-based applications. Linked Data and Semantic Web technologies make it easy to connect heterogeneous datasets without advance coordination between data producers and consumers.

The TWC LOGD Portal's landing page by itself mashes up data from multiple sources. As shown in Figure 3, the content panels are based on live SPARQL queries over several data sources and the query results are rendered using XSLT and Google APIs. The metadata about demos/tutorials on the Portal is maintained in a Semantic Drupal [12] based infrastructure and published in RDFa-annotated XHTML pages. The semantic annotations are loaded into our LOD Cache (`http://data-gov.tw.rpi.edu/ws/lodcx.php`), and the "demo" panel renders the corresponding SPARQL query results using a dedicated XSLT style sheet. The metadata of the converted LOGD datasets is automatically loaded into the TWC LOGD's SPARQL endpoint upon the completion of LOGD dataset conversion, and the "LOGD Stats" panel issues a couple of SPARQL queries to count the number of RDF triples and OGD datasets hosted on the Portal. We also leverage the Google Feed API[34] to integrate several RSS feeds, which are generated using unique technologies from different sources: the TWC LOGD Website uses native Drupal functions to generate RSS feeds for recently updated tutorials and videos; the Data-gov Wiki maintains TWC-LOGD-relevant news in RDF and the corresponding SPARQL query results are turned into RSS feeds using a user-contributed web service from Yahoo! Pipes[35]; RSS Feeds for other Data.gov-relevant news are generated from a Google News search; and a SemDiff [4] service is used to compute RSS feeds showing the recent updates in the Data.gov dataset catalog.

The TWC LOGD Portal is not merely one application but an ecosystem promoting the integration of conventional web technologies and LOGD innovations for consuming government data. Over 50 mashups and visualizations have been created to demonstrate the diverse application of LOGD, including: (i) integrating data from multiple sources such as DBpedia, the New York Times Data API[36], Twitter and OGD produced by both US and non-US sources; (ii) deploying LOGD data via web and mobile interfaces (e.g., mobile version of the White House Visitor Search application has been released in the Apple Store[37]); (iii) supporting interactive analysis in specific domains including health [13], policy [14] and financial data [15]; (iv) consuming readily-available Web-based services (e.g., Yahoo! Pipes[38], IBM ManyEyes[39], and Microsoft Web N-gram service [16]); and (v) building semantic data access and integration tools (e.g., semantic search[40] and Data.gov dataset search[41]).

4.3. Achievements and Community Impact

The TWC LOGD Portal demonstrates that Linked Data and Semantic Web technologies can be effectively applied to reduce development costs, promote collaborative data integration, and increase the reuse of data models, data links and visualization techniques in the government domain.

Our work on LOGD continues to demonstrate that developers don't need to be experts in semantic tech-

[30] `http://www.mindswap.org/~kendall/sparql-results-json`
[31] `http://logd.tw.rpi.edu/ws/sparqlproxy.php`
[32] `http://code.google.com/apis/visualization`
[33] `http://jquery.com`
[34] `http://code.google.com/apis/feed`
[35] `http://pipes.yahoo.com/pipes/pipe.info?_id=4599dc7d331b04f0f9cefa6529cf8280`
[36] `http://developer.nytimes.com`
[37] `http://itunes.apple.com/us/app/twc-white-house-visitors/id399556322`
[38] `http://pipes.yahoo.com`
[39] `http://www-958.ibm.com`
[40] `http://data-gov.tw.rpi.edu/ws/lodcs.php`
[41] `http://news.rpi.edu/update.do?artcenterkey=2804`

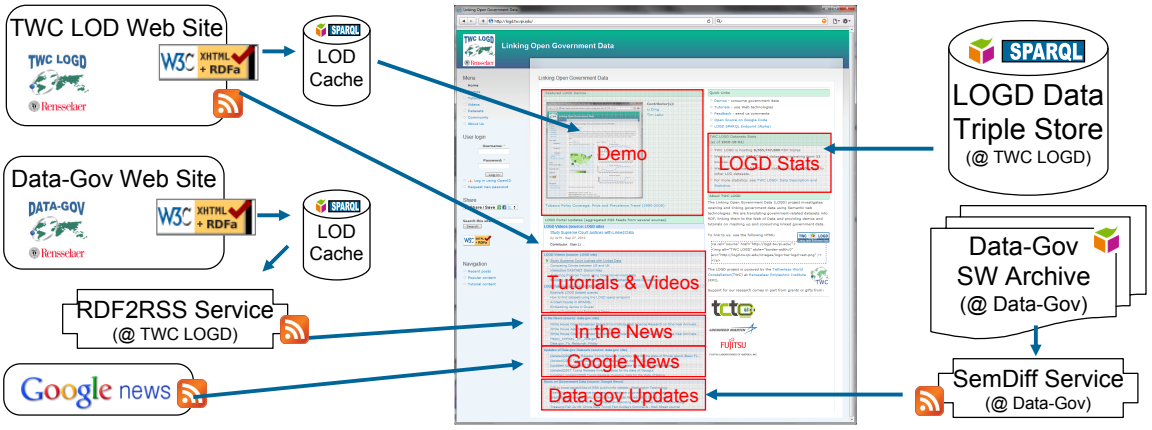

Figure 3: The TWC LOGD Portal's landing page as a dynamically-sourced mashup

nologies or Linked Data to create useful, semantically-enabled LOGD applications. In particular, undergraduate students in RPI's 2009 and 2010 Web Science classes created mashups as course projects using tools and SPARQL endpoints from the TWC LOGD Portal. Given a two-hour introduction to the basics about RDF and SPARQL and examples on using visualization tools, each student group was able to create visualizations based on at least two LOGD datasets within two weeks. In August 2010, the Data.gov project hosted a Mash-a-thon workshop[42], taught by graduate students from RPI, to engage government developers and data curators in hands-on learning using tools and datasets from the Portal. 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 Linked Data best practices embodied by the Portal.

5. LOGD Community

Having demonstrated the specific community contributions made by the TWC LOGD Portal in the areas of LOGD production and consumption, we now discuss the collaboration mechanisms enabled by the Portal.

5.1. Transparency and Provenance in Collaboration

The distributed nature of LOGD ecosystems raises concerns about the integrity of the resulting data products. In the TWC LOGD Portal, transparency is especially critical since it involves an academic research center aggregating governmental datasets from both government and non-government sources, each with its own degree of authority, policies, documentation, and trustworthiness. Provenance traces are captured in the Portal for both LOGD production and LOGD consumption, enabling data consumers to debug mashups collaboratively [17], provide a stronger basis for trust in mashup results [18], and potentially provide explanations for the workflows behind mashups [19]; this is an area of exploration for the TWC LOGD team.

The provenance trace of a LOGD production workflow is automatically recorded by `csv2rdf4lod` using PML, and the provenance trace of LOGD consumption workflow is manually asserted on the Portal's RDFa-yielding Drupal pages. The value of provenance has also been exhibited through LOGD mashups. In August 2010, we created a map-based visualization to compare a number of variables including smoking prevalence, tobacco policy coverage, tobacco tax and prices[43]. The first version of this visualization put the policy coverage data on the map with a nominal range of 0% to 100%. An end user discovered that the maximum value reported was actually 101%. Further investigation with our government contact confirmed that this observation was the result of a rounding issue when adding up source data entries. The issue was quickly resolved in the data publishing process and the anomaly was removed. In another case, a user reported that one of our demos used a dataset that contained only a portion of another, more comprehensive, dataset that had been released after the demo was originally created; we were able to update the mashup by simply changing the dataset information in its main query. Through these

[42] `http://www.data.gov/communities/node/116/view/119`

[43] `http://logd.tw.rpi.edu/node/3860`

examples we see that the exposed provenance traces enable effective communication between users and government data curators.

5.2. Scalability enabled by Community Participation

The scalability of the TWC LOGD Portal's infrastructure has in part been demonstrated by our success in downloading, converting and publishing a large and diverse collection of OGD datasets on an ongoing basis. The necessary human workload (e.g., data cleansing and linking) are distributed to the LOGD community: a pool of undergraduate students can help clean up OGD raw data into CSV tables while our advanced graduate research assistants can focus on building supporting *tools*, such as `csv2rdf4lod`, `sparqlproxy` and Semantic Web extensions for `Drupal`. To grow the LOGD community, we have developed a rich set of *educational resources*, including replicable open source demos and tutorials that teach web developers cutting-edge technologies as well as best practices for building their own LOGD mashups. We also regularly use these online instructional materials as the basis for our ongoing LOGD community development activities including face-to-face hack-a-thons, courses and workshops.

6. Related Work

Open government data initiatives typically originate with the publication of online catalogs of raw datasets; these catalogs typically feature keyword search and faceted browsing interfaces to help users find relevant datasets and retrieve the corresponding metadata including dataset description and download URLs. For example, Data.gov maintains three dataset catalogs including the *Raw Data Catalog*, the *Tool Catalog* and the *Geodata Catalog*: the first two catalogs share a Socrata-based faceted search interface[44], while the *Geodata Catalog* provides a separate interface[45]. The OpenPSI Project[46] collects RDF-based catalog information about the UK's government datasets to support government-based information publishers, research communities, and web developers. CKAN (Comprehensive Knowledge Archive Network)[47] is an online registry for finding, sharing and reusing datasets. As of January 2011 about 1600 datasets have been registered on CKAN, and this collection has been used to generate the LOD cloud diagram and support dataset listings in Data.gov.uk. CKAN's dataset metadata is natively published in JSON format, but it is also experimenting with RDF encoding[48] using DERI's Data Catalog Vocabulary (dcat)[49]. The TWC LOGD Portal has collected 50 dataset catalogs (including the three Data.gov catalogs), covering 323,304 datasets from 18 countries and two international organizations, and our current work explores a faceted search interface for integrating international dataset catalogs[50].

Instead of providing dump files, several OGD projects offer Web-based data APIs to expose government datasets to web applications. For example, the Sunlight Foundation[51] has created the National Data Catalog[52] which makes federal, state and local government datasets accessible via a RESTful data API. Socrata is a Web platform for publishing datasets that provides a full catalog of all their open government datasets, along with tools to browse and visualize data, and a RESTful data API for developers. Microsoft has also entered this space with their OData data access protocol and their Open Government Data Initiative (OGDI)[53]; recently a small number of OGD datasets have been published on Microsoft's Azure Marketplace DataMarket[54]. Currently, these platforms are not interlinked. None of their data APIs provide a means for developers to see or reuse the underlying data model, making it hard to extend existing data APIs or mashing up data from multiple data APIs.

There are an increasing number of Linked Data-based projects involving government data in the US and around the world. GovTrack[55] is a civic project that collects data about the U.S. Congress and republishes the data in XML and as Linked Data. Goodwin et al. [20] used linked geographical data to enhance spatial queries on the administrative geographic entities in Great Britain. Data.gov.uk, the official open government data portal of the UK, has released LOGD datasets together with OGD raw datasets since its launch in January 2010. As we discussed earlier, the TWC LOGD Portal not only produces and publishes LOGD but also provides open source tools and educational resources to help others (including Data.gov) participate in collaborative LOGD production.

[44] `http://explore.data.gov/browse`
[45] `http://www.data.gov/catalog/geodata`
[46] `http://www.openpsi.org`
[47] `http://ckan.net`
[48] `http://semantic.ckan.net`
[49] `http://vocab.deri.ie/dcat`
[50] `http://logd.tw.rpi.edu/node/9903`
[51] `http://sunlightfoundation.com`
[52] `http://nationaldatacatalog.com`
[53] `http://ogdi.codeplex.com`
[54] `https://datamarket.azure.com`
[55] `http://www.govtrack.us`

7. Conclusions

The TWC LOGD Portal demonstrates a model infrastructure and several workflows for linked open government data deployment. Our experiences have shown the valuable role Linked Data and Semantic Web technologies can make in the open government data domain, and these efforts have been recognized by the US government data-sharing community. The Portal has also served as an important training resource as these technologies have been adopted by Data.gov, the US federal open government data site.

The TWC LOGD Portal has done much to foster a LOGD community but there are many improvements that can be made moving forward. 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 site. The recent TWC-led government Mash-a-thon highlighted the value of interaction between the participants with the TWC LOGD team; we hope to position the Portal as a 24x7x365, community-driven extension of that interaction model.

The publication of converted government datasets is a critical service provided by the TWC LOGD Portal and thus we plan to significantly extend the scope of our LOGD dataset production. In particular, TWC is exploring how best to add the nearly 300,000 US government geodata datasets to the TWC LOGD conversion workflow and we are working on demos and tutorials to facilitate consumption and reuse of this new class of data. Data sources other than Data.gov are being included in the Portal. Further, we are working on more extensive encoding and exposure of provenance information in addition to creating tools that enable user-contributed annotation of demos, giving users the ability to identify potential data issues or updates.

The LOGD world is vast and growing exponentially. It must provide services to a diverse set of stakeholders ranging from providers, curators, and developers, to civil servants, activists, media, community leaders, and average citizens. The long-term goal is for the TWC LOGD Portal to become one focal point for engaged discussion and outreach centered on LOGD issues, technologies and best practices, as well as to help create communities of citizens who can help create new ways of interacting with their governments.

References

[1] A. Smith, Government online, URL: `http://www.pewinternet.org/Reports/2010/Government-Online.aspx`, accessed on Jan 25, 2011 (2010).

[2] T. Berners-Lee, Putting government data online, `http://www.w3.org/DesignIssues/GovData.html`, accessed Sep 25, 2010 (2009).

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

[4] L. Ding, D. Difranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, J. Hendler, Data-gov wiki: Towards linking government data, in: Proceedings of the AAAI 2010 Spring Symposium on Linked Data Meets Artificial Intelligence, 2010.

[5] C. Bizer, T. Heath, T. Berners-Lee, Linked data - the story so far, International Journal on Semantic Web and Information Systems 5 (3) (2009) 1–22.

[6] L. Ding, D. Difranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinness, J. Hendler, Twc data-gov corpus: Incrementally generating linked government data from data.gov, in: WWW'10 (developer track), 2010, pp. 1383–1386.

[7] T. Lebo, G. T. Williams, Converting governmental datasets into linked data, in: Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, 2010, pp. 38:1–38:3.

[8] D. L. McGuinness, L. Ding, P. P. da silva, C. Chang, Pml 2: A modular explanation interlingua, in: Proceedings of the AAAI 2007 Workshop on Explanation-Aware Computing, 2007.

[9] J. Flores, L. Ding, Discovering the hidden cross-dataset links in data.gov, in: Web Science'11, 2010.

[10] M. Krötzsch, D. Vrandecic, M. Völkel, Semantic mediawiki, in: ISWC'06, 2006, pp. 935–942.

[11] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, D. Sheets, Tabulator: Exploring and analyzing linked data on the semantic web, in: Proceedings of the 3rd International Semantic Web User Interaction, 2006.

[12] S. Corlosquet, R. Delbru, T. Clark, A. Polleres, S. Decker, Produce and consume linked data with drupal!, in: ISWC'09, 2009, pp. 763–778.

[13] P. Courtney, A. R. Shaikh, N. Contractor, D. McGuinness, L. Ding, E. M. Augustson, K. Blake, G. D. Morgan, R. Moser, G. Willis, B. W. Hesse, Consumer health portal: An informatics tool for translation and visualization of complex, evidence-based population health data for cancer prevention and control, in: Proceedings of the 138th APHA Annual Meeting, 2010.

[14] X. Li, L. Ding, J. A. Hendler, Study supreme court decision making with linked data, in: Web Science'10, 2010.

[15] X. Li, J. Bao, J. A. Hendler, Fundamental analysis powered by semantic web, in: Proceedings of IEEE Symposium on Computational Intelligence for Financial Engineering and Economics, 2011.

[16] J. Huang, J. Gao, J. Miao, X. Li, K. Wang, F. Behr, C. L. Giles, Exploring web scale language models for search query processing, in: WWW '10, 2010, pp. 451–460.

[17] J. Michaelis, D. L. McGuinness, Towards provenance aware comment tracking for web applications, in: The Third International Provenance and Annotation Workshop (IPAW 2010), 2010, pp. 265–273.

[18] X. Li, T. Lebo, D. L. McGuinness, Provenance-based strategies to develop trust in semantic web applications, in: The Third International Provenance and Annotation Workshop (IPAW 2010), 2010, pp. 182–197.

[19] D. L. McGuinness, V. Furtado, P. P. da Silva, L. Ding, A. Glass, C. Chang, Explaining semantic web applications, in: Semantic Web Engineering in the Knowledge Society, Information Science Reference, 2008, pp. 1–24, (chapter 1).

[20] J. Goodwin, C. Dolbear, G. Hart, Geographical linked data: The administrative geography of great britain on the semantic web, Transactions in GIS 12 (s1) (2009) 19–30.