Self-Service Linked Government Data with dcat and Gridworks

Richard Cyganiak
Digital Enterprise Research Institute, NUI Galway, Ireland
richard@cyganiak.de

Fadi Maali
Digital Enterprise Research Institute, NUI Galway, Ireland
fadi.maali@deri.org

Vassilios Peristeras
Digital Enterprise Research Institute, NUI Galway, Ireland
vassilios.peristeras@deri.org

ABSTRACT

Open Government Data initiatives in the U.S., U.K. and elsewhere have made large amounts of raw data available to the public on the Web. There is enormous potential in applying Linked Data principles to these datasets. This potential currently remains largely untapped because governments lack the resources required to convert from raw data to high-quality Linked Data on a large scale. We present a "self-service" approach to this problem: By connecting a powerful Gridworks-based data workbench application directly to data catalogs, via a standard Data Catalog Vocabulary, data professionals outside of government can contribute to the Linked Data conversion process, thus obtaining data for their own needs and benefiting the larger Linked Government Data effort.

1. INTRODUCTION

Open Government is a recent movement towards more openness and transparency in government. *Open data* is an important part of this movement. Governments produce huge amounts of information as part of their daily operations, ranging from statistics used in policy-making to internal financial auditing records. Opening up this information—whose creation has ultimately been paid by the taxpayer—increases transparency and accountability, provides businesses and citizens with valuable information, enables third parties to provision new services based on government data, and removes intra-government red tape. Open government is becoming an official policy in several jurisdictions, especially in the U.S., U.K., and in Australia. It will increasingly spread throughout the European Union as EC directives get translated into local law.

1.1 Data catalogs

One cornerstone of an open government policy is the *data catalog*, exemplified by the US' *data.gov* and the UK's *data.gov.uk* websites. They serve as central one-stop portals where interested publics can find data published by government bodies, and are important for giving visibility to the process of translating policy into reality. More than 40 government data catalogs are now in operation, listing national (US, UK, Australia, New Zealand), regional (e.g., Kent County, Basque Country), local (e.g., London, New York, Vancouver), or thematic (e.g., geospatial, statistical) datasets.

1.2 Linked Government Data (LGD)

Data catalogs support discovery of available information, but working with this information can still a challenge. Often it is provided in a haphazard way, driven by practicalities within the producing government agency, and not by the needs of the information user. Formats are often inconvenient, (e.g., numerical tables as PDFs), there is little consistency across datasets, and documentation is often poor.

Linked Government Data is a promising technique to lessen these challenges. The benefist of applying Linked principles to government data are (i) simpler access through a unified data model; (ii) good support for documenting data semantics; (iii) re-use of vocabularies; (iv) allows fine-grained referencing of information through the use of URIs; (v) allows interlinking of related information, making it possible to automatically pull together all information about a school, ward, or other entity.

But realizing this potential is costly. The pioneering Linked Government Data efforts in the U.S. and U.K. have shown that creating high-quality Linked Data from raw data files requires considerable investment into reverse-engineering, documenting data elements, data cleanup, schema mapping, and instance matching. At *data.gov*, large numbers of datasets were converted to RDF using a simple automatic algorithm, without much curation effort, which limits the practical value of the resulting RDF. In the U.K., RDF datasets published around *data.gov.uk* are carefully curated and of high quality, but due to limited availability of trained staff and contractors, only selected high-value datasets have been subjected to the Linked Data treatment, while most data remains in raw form.

1.3 Self-service LGD

We present a contribution towards solving this cost problem of Linked Government Data: the "self-service approach". It shifts the burden of Linked Data conversion towards the data consumer. This has several advantages: (i) There are more of them; (ii) they are reasonably likely to have both the necessary skills and the motivation for performing conversion

and cleanup; (iii) they know which datasets they need, and don't have to rely on the government's data team to convert the right datasets. To achieve this "self-service approach", we have created three components:

1. A standard RDF Schema vocabulary for expressing a data catalog as Linked Data.
2. A Linked Data site and SPARQL endpoint that republishes the contents of four major government data catalogs using this vocabulary.
3. A desktop application that allows users to connect to data catalogs, browse the catalog, open tabular datasets (such as CSV and Excel files), perform cleanup and consolidation, map them to arbitrary RDF vocabularies, and export them as RDF. This desktop application is an extension of Freebase Gridworks.

A fourth component is still missing at this time: a mechanism for contributing the results of cleanup and conversion back to a central place, where other potential users of the same dataset can find it without having to repeat the conversion effort. Currently, users would have to rely on out-of-band solutions, such as the ckan.net website or dedicated Linked Government Data mailing lists, along with their own web hosting, to share and announce their conversion results.

2. THE DATA CATALOG VOCABULARY

Data catalogs are websites. Users access them through their web browsers and use the site functionality that the catalog operator has put into place. This model is limiting. Automated processing that goes beyond what's provided by the catalog operator is only possible if the catalog's contents to also be available in a machine-readable form. Examples include advanced querying and filtering, bulk format conversion, or aggregation of several catalogs into a super-catalog, such as the Guardian's World Government Data site or the Sunlight Foundation's National Data Catalog.

Currently, only a few catalogs provide their contents in machine-readable form. Those who do, have adopted different formats (CSV, Atom, RDFa) and schemas.

To overcome this problem, we have proposed a standard RDF Schema vocabulary for data catalogs, called *dcat*. We developed *dcat* based on a survey of the metadata schemas of seven different data catalogs. The resulting vocabulary makes heavy use of existing vocabularies such as Dublin Core, SKOS and FOAF. To involve a larger number of stakeholders, we have recently moved *dcat* development to a Task Force within the W3C's eGovernment Interest Group. More information is available at `http://www.w3.org/egov/wiki/Data_Catalog_Vocabulary`.

3. BRINGING DATA CATALOGS INTO THE WEB OF LINKED DATA

To put data catalog interoperability into practice, we selected four data catalogs (the national catalogs of the US, UK, Australia and New Zealand) and published their contents as Linked Data in *dcat* format.

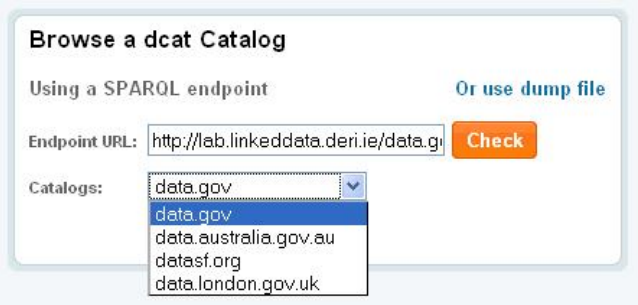

Figure 1: Connecting to a *dcat*-enabled catalog

Depending on the catalog, the raw data was obtained by CSV download, (non-*dcat*) RDFa extraction, or screen-scraping. We stored the catalog contents in a relational database, mapped the metadata model of each catalog was mapped to the *dcat* vocabulary, and published as Linked Data and as a SPARQL endpoint using D2R Server. This effort has made descriptions of 4,500 public sector information datasets, with a total of 60,000 triples, available as Linked Data using a uniform vocabulary. The data can be accessed at `http://lab.linkeddata.deri.ie/govcat`, and more information including example queries is available at `http://vocab.deri.ie/dcat-overview`.

Our ultimate goal is to get catalog operators to publish *dcat* directly on their catalog websites. We are working with several of them towards this goal in the context of the W3C eGovernment IG. The *data.gov.uk* team has begun deploying *dcat* as embedded RDFa markup on their website.

4. EXTENDING FREEBASE GRIDWORKS

Freebase Gridworks is an open-source desktop application for exploring, manipulating and converting tabular datasets, originally developed by Metaweb. Data is presented in a tabular grid. For any column, facets and value distribution diagrams can be created, in order to filter and explore the data. Various functions oriented towards cleanup and value consolidation can be applied to columns, including the powerful *Gridworks Expression Language (GEL)*. The dataset can be mapped to the Freebase schema, and exported into Freebase, a CSV file, or Excel file. More information about Gridworks, including screencasts, can be found at `http://code.google.com/p/freebase-gridworks/`.

We have extended Gridworks in various ways. First, we have added the ability to connect to *dcat*-enabled data catalogs (Fig. 1). Once connected, the dataset list can be browsed via a faceted interface and text search over the dataset metadata (Fig. 2). Any datasets in tabular formats (CSV, Excel) can then be opened in the standard Gridworks UI.

We note that the functionality described up to here has clear benefits to non-Linked Data users: They can quickly connect to and browse the contents of data catalogs, find relevant raw datasets, clean them up, and export them to Excel or CSV for further processing using Gridworks' pre-existing export features.

However to achieve high-quality RDF export, we have added

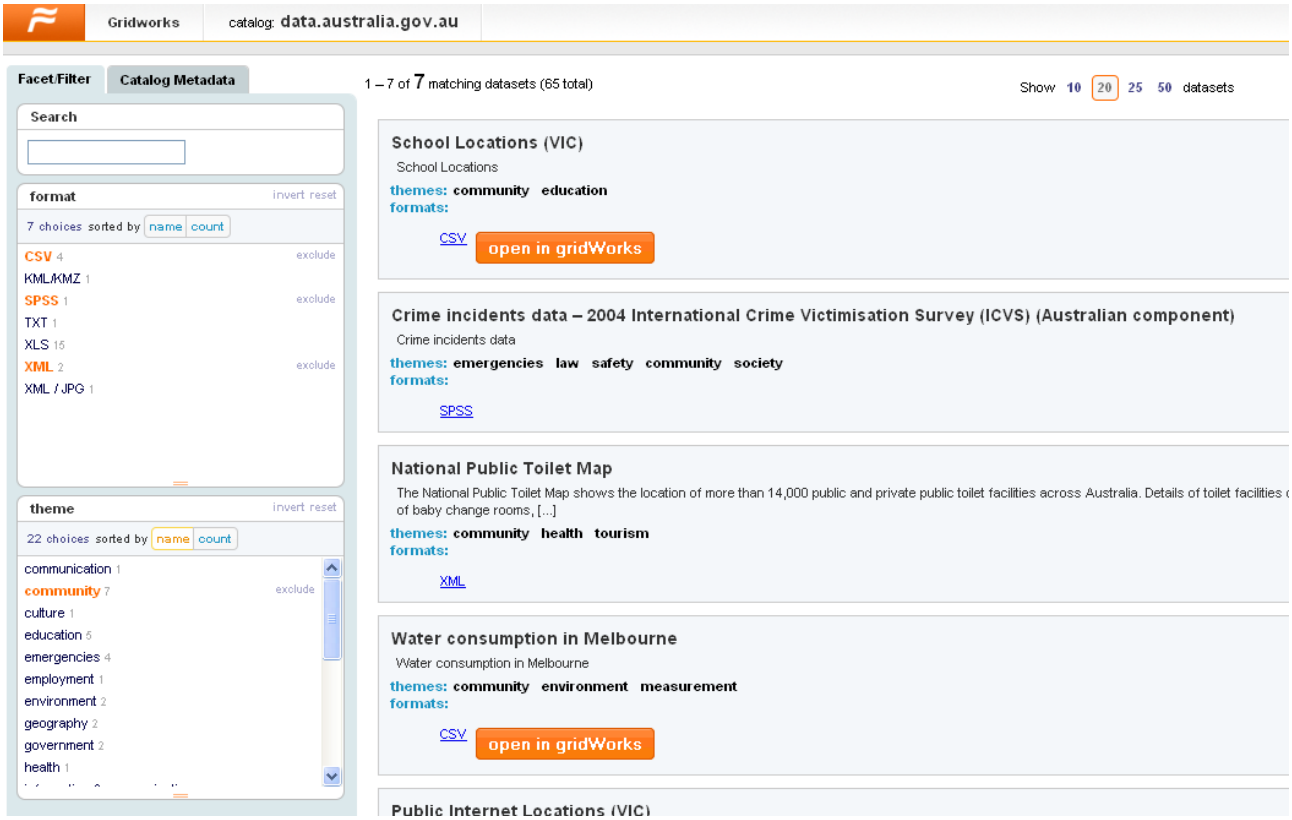

Figure 2: Browsing a data catalog

an RDF Schema mapping facility to Gridworks. It is based on the Freebase schema mapper. The user defines a tree-shaped pattern and annotates each edge with an RDF property. For each node, he or she specifies how it ought to be created from a cell value, and wether a blank node, URI or literal ought to be created. Optionally, an RDF class can be specified as well. Selection of classes and properties is assisted by autocompletion. A list of popular vocabularies from *prefix.cc* is preloaded; additional terms can be added. RDF is generated by instantiating the tree pattern for each row in the table (Fig. 3).

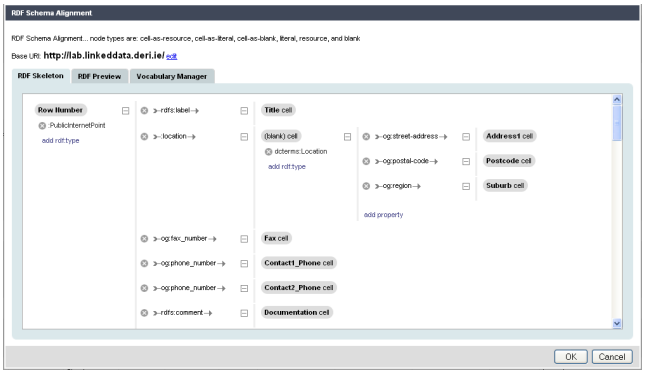

Figure 3: Mapping a table to an RDF schema

For downloads and documentation, please refer to the project website at `http://lab.linkeddata.deri.ie/2010/dcat`.

5. DISCUSSION AND CONCLUSION

We have shown an application that allows users to browse government data catalogs, open tabular datasets listed in the catalogs, explore and clean up and structure the datasets using Gridworks' powerful user interface, and map the result to existing RDF vocabularies for RDF export. This empowers Linked Data users to contribute to the process of engineering raw government data into high-quality Linked Data. We envision that users should be able to contribute the cleanup steps, RDF mappings, and resulting triples back to a central place, for example by connecting to a SPARQL UPDATE supporting endpoint, but this is not yet implemented.

Datasets can be interlinked either by manually adding URIs to the dataset, or by using the Gridworks Expression Language to express rules for generating URIs from existing cell data. Gridworks also contains a sophisticated reconciliation engine that can link cells to Freebase topics; a similar mechanism should be available for linking to existing Linked Datasets and government-defined URI sets.

We have just addressed tabular data. Other types of data, such as geospatial information, event feeds, and complex statistical data, require similar solutions with specialised user interfaces geared towards the type of data. Imagine connecting to a *dcat*-enabled data catalog from within Google Earth or professional GIS packages.