TWC International Open Government Dataset Catalog *

John S. Erickson†
erickj4@rpi.edu

Yongmei Shi
sym@cs.rpi.edu

Li Ding
dingl@cs.rpi.edu

Eric Rozell
rozele@rpi.edu

Jin Zheng
zhengj3@rpi.edu

James A. Hendler
hendler@cs.rpi.edu

Tetherless World Constellation
Rensselaer Polytechnic
Institute
110 8th Street
Troy, NY 12180

ABSTRACT

The TWC International Open Government Dataset Catalog (IOGDC) integrates a diverse selection of more than 70 government dataset catalogs from around the world. IOGDC demonstrates a practical dataset catalog metadata model for integrating diverse dataset catalogs collected from the real world and linking those catalogs into Linked Data Cloud. IOGDC's faceted browsing and search interface provides a scalable and reconfigurable solution for finding and browsing open government datasets which also offers a compelling demonstration of the value of a common metadata model for open government dataset catalogs. We believe that the vocabulary choices demonstrated by IOGDC highlight the potential for useful Linked Data applications to be created from open government catalogs and will encourage the adoption of such a standard worldwide.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Semantic Web*

General Terms

standardization

Keywords

metadata, Linked Data, open government, government data

*A submission to the *Open Government Data* track of the I-SEMANTICS 2011 Triplification Challenge

†Corresponding author

I-SEMANTICS 2011 Graz, Austria

1. INTRODUCTION

The TWC International Open Government Dataset Catalog (IOGDC) is a data integration project for harvesting, cleaning, aggregating, and republishing the metadata catalogs of open government datasets from around the world. As of July 2011 more than 70 catalogs from over 25 countries have been harvested by IOGDC, linking to more than 335,000 datasets from a wide variety of categories published online by different levels of government including local, state, provincial, federal and others. The catalog data is republished as Linked Data via de-referencable URIs (following Linked Data principles) and via gzipped dump files[1]. A scalable, web-based faceted browsing and search interface [2] has been developed to help users find datasets using the republished catalog data. TWC IOGDC's success at exposing existing, proprietary dataset catalogs as Linked Data is illustrated by the following:

IOGDC integrates a diverse selection of government dataset catalogs. While a number of open government dataset catalogs, including Data.gov, Data.gov.uk, United Nations and OpenEI.org have already provided dataset registration and search capabilities, they are usually maintained by independent contributors and their cataloged datasets are typically limited to specific regions or domains. IOGDC demonstrates an incremental process that integrates semi-automatic dataset catalog harvesting activities and automated Linked Data production and publishing.

IOGDC demonstrates a practical dataset catalog metadata model for integrating diverse dataset catalogs collected from the real world and linking the catalogs into Linked Data Cloud. In 2011 the W3C eGovernment Interest Group (eGov IG) [3] began developing the Data Catalog (DCAT) model. While the vocabulary choices made for IOGDC (see Appendix) do not precisely match DCAT, they are close enough to demonstrate the potential of the DCAT model. Like DCAT, IOGDC uses existing vocabularies and alignment ontologies where possible,

[1] `http://logd.tw.rpi.edu/2011/iogdc-dump-all-v20110608.tar.gz`
[2] `http://logd.tw.rpi.edu/demo/international_dataset_catalog_search`
[3] http://www.w3.org/egov/

including datasets from the Linked Open Data cloud.

IOGDC's faceted browsing and search interface provides a scalable and reconfigurable solution for finding and browsing open government datasets. The IOGDC's user interface is an efficient faceted browser that illustrates the value of integrated open government dataset catalogs and the supporting Semantic Web technologies.

2. OGD CATALOG COLLECTION

OGD dataset catalogs are collected for IOGDC through a semi-automated process that includes manual catalog identification and customized semi-automated catalog harvesting. The resulting Linked Data may be consumed by external applications and services through SPARQL endpoints and RDF downloads or directly by users through an efficient faceted browser.

Catalog Identification. Government data catalogs are released on the Web in a variety of approaches. In order to locate the web page that publishes catalogs, we use manual Google search and well-known "catalogs of catalogs" from e.g., Open Knowledge Foundataion's catalog [4]. Currently, IOGDC has collected more than 70 catalogs from national, regional, state and city governments around the world.

Catalog Harvesting. An identified catalog might have been published using one of many different approaches. RDF-native catalogs are extremely rare; more common are RDBMS-based listings of datasets formatted as ASP or JSP-generated HTML tables or as simple lists. AJAX-based catalogs are more complex since the harvesting code must deal with dynamic page content which is intended to be interpreted and executed by the client-side browser. The TWC team targets catalogs in which the structure is evident, the local metadata is clear, the content are fairly static, and where the country or region has been under-represented in IOGDC. To further reduce complexity and especially to enable undergraduate assistants to work on catalog harvesting, we limit the output of initial harvesting output to CSV files in restricted structure (e.g. including header row and several key columns such as dataset tile and homepage). Most IOGDC scraper code can only be re-used when targeted catalogs share *exactly* the same structure. e.g. Oregon [5], Baltimore[6], and Washington, DC [7]. However, the catalog harvesting tasks are highly independent, therefore, many undergraduate assistants can work in parallel in "course project" style, submitting their scraper code together with their results using Subversion [8]; this model also enables reuse code from existing submitted scrapers if desired.

3. LINKED DATA PRODUCTION

Although most catalogs share certain common dataset properties including title, description, homepage, and download URLs, the metadata from different catalogs usually adopt different metadata structures and vocabularies. In order to integrate over 70 catalogs and make them available as Linked Data, we used the following strategies.

Linking metadata structure. We developed a metadata model to describe *catalog* and *dataset* concepts (see Appendix). Based on an earlier proposal for government dataset description[9] and informed by DCAT, our vocabulary was developed to fit the existing dataset catalogs found in the wild. For example, two dataset catalogs (i.e., Raw Data catalog and Tool data catalog) are released from one CSV file from Data.gov, leading us to introduce two-part catalog title structure using *dgtwc:catalog_title* and *dgtwc:catalog_subtitle*.

All catalogs are serialized in CSV before being converted into Linked Data, regardless of whether they were generated by harvesting scrapers or manual process. Based on a automatically generated and manually edited conversion configuration, the CSV files are then converted to RDF using TWC's LOGD pipeline, built around the `csv2rdf4lod` conversion and enhancement tool.[10] Manual edits on the conversion configurations enable two important types of links: (i) mapping divers column header names into our universal metadata vocabulary to enable RDF property level linking; and (ii) use DBpedia URIs in the value of properties, e.g. use *dbpedia:United_States* as the value of *dgtwc:catalog_country*. The result Linked Data is currently accessible through our SPARQL endpoint. [11] as well as dataset dumps.

4. FACETED BROWSING AND SEARCH

The value of an integrated international dataset catalog is demonstrated by the International Dataset Catalog Search application. Figure 1 illustrates the faceted browsing interface of a search for "energy" data. The demo is an AJAX system, where the client query and access catalog metadata via a RESTful web service interface that is compatible with the OpenSearch protocol [12]. This design allows us to hide the triplestore implementation on server side from the front end faceted browsing component. On the service side, we employ OpenLink Virtuoso features to provide efficient full-text search (`bif:contains`) rather than using regular expressions offered by SPARQL standard. On the front end side, we use S2S [13] and leverage CSS to control layout. The S2S faceted browser was developed as part of the Semantic eScience Framework project (SeSF) [14] was adapted to provide a highly efficient faceted browse and search experience for the user.

5. CONCLUSION AND FUTURE WORK

Public demonstrations of IOGDC began in May 2011 and have resulted in constructive discussions with the CKAN and W3C eGov communities regarding the effectiveness of our catalog and dataset metadata models. While a lot of properties have been defined in DCAT, CKAN and our metadata model, limited metadata can be obtained from real world dataset catalogs: even the popular "category" and "homepage" properties are not necessarily provided in all catalogs. Therefore, the catalog metadata model should be revisited in the future to better adpate reality.

[4] `http://opengovernmentdata.org/data/catalogues/`
[5] `http://data.oregon.gov/`
[6] `http://data.baltimorecity.gov/`
[7] `http://data.dc.gov/`
[8] `http://subversion.apache.org/`
[9] `http://data-gov.tw.rpi.edu/wiki/TWC_Data-gov_Vocabulary_Proposal`
[10] `https://github.com/timrdf/csv2rdf4lod-automation/wiki`
[11] `http://logd.tw.rpi.edu/sparql`
[12] `http://www.opensearch.org/`
[13] `http://tw.rpi.edu/web/project/sesf/workinggroups/s2s`
[14] http://tw.rpi.edu/web/project/SeSF

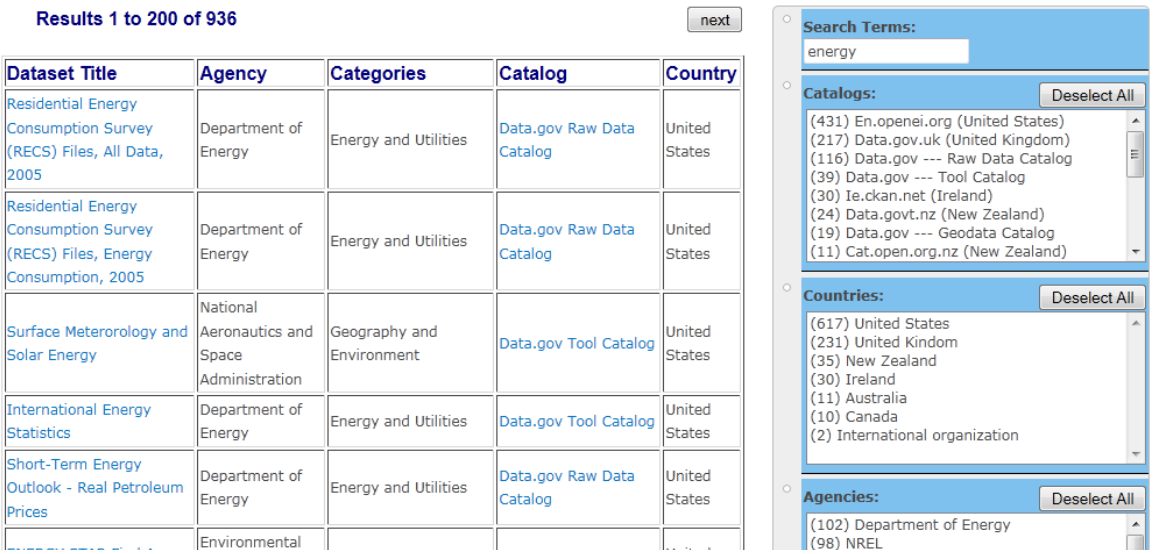

Figure 1: Faceted browsing based International Government Dataset Search

TWC's strategy of scraping, converting and publishing international open dataset catalog metadata has been useful in the short term, but it is unlikely to scale and has great potential for error. Ultimately, our methods should not be necessary; governments should be their own authoritative sources for catalog data. Our fervent hope is that IOGDC will demonstrate to governments and other open data providers worldwide the value of publishing catalog metadata as linked open data using a standard metadata model combined with a catalog metadata harvesting protocol.

APPENDIX

A. TWC IOGDC METADATA MODEL

Namespaces Used

dgtwc: http://data-gov.tw.rpi.edu/2009/data-gov-twc.rdf#
dcterms: http://purl.org/dc/terms/
foaf: http://xmlns.com/foaf/0.1/
conversion: http://purl.org/twc/vocab/conversion/

properties of *conversion:DatasetCatalog*

- **dcterms:title** (Range: rdfs:Literal): Title or Name of the catalog
- **foaf:homepage** (Range: rdfs:Resource): URL of web page or unique identifier for accessing catalog
- **dcterms:description** (Range: rdfs:Literal): Description of the catalog
- **dgtwc:number_of_datasets** (Range: xsd:Integer): Estimated number of datasets listed in catalog
- **dcterms:spatial** (Range: rdfs:Resource): Spatial Region. Primary geographical region covered by catalog datasets
- **dgtwc:spatial_granularity** (Range: rdfs:Resource): Spatial Region Type. Use following values: Worldwide, National, State, City, ...

properties of *conversion:CatalogedDataset*

- **dcterms:title** (Range: rdfs:Literal): Title or Name of the dataset
- **foaf:homepage** (Range: rdfs:Resource): URL of web page or unique identifier for accessing dataset
- **dcterms:description** (Range: rdfs:Literal): Description of the dataset
- **dcterms:identifier** (Range: rdfs:Literal): Unique string ID of the dataset
- **dcterms:download_format** (Range: rdfs:Literal): Available formats of the dataset
- **dgtwc:category** (Range: rdfs:Literal): Original category (asserted by the catalog) of dataset
- **dgtwc:categories** (Range: rdfs:Literal): All categories (asserted by the catalog) of this dataset
- **dcterms:subject** (Range: rdfs:Literal): Original keywords (asserted by the catalog) of this dataset
- **dgtwc:keywords** (Range: rdfs:Literal): All keywords (or tags) of this dataset
- **dgtwc:agency** (Range: rdfs:Resource): Actual publisher (typically government agency) who curates, publishes the dataset
- **dgtwc:catalog_title** (Range: rdfs:Literal): Government website (Country/State/City/Organization) which owns dataset catalog listing dataset
- **dgtwc:catalog_homepage** (Range: rdfs:Resource): Catalog URL: URL of web page or unique identifier of dataset catalog listing dataset.
- **dgtwc:catalog_subtitle** (Range: rdfs:Literal): Used for Data.gov's three sub-catalogs: Raw Data Catalog, Tool Catalog, and GeoData Catalog
- **dgtwc:catalog_country** (Range: rdfs:Literal): Country of owner of dataset catalog; strong indicator of spatial region covered by listed datasets.