First they have to **find** it: Getting Government Data Discovered and Used

John S. Erickson, Ph.D.
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, New York, USA

Twitter: @olyerickson  #TWCRPI
Starting with efforts in the US and UK, governments around the world have recognized the need to publish their critical data.
Diverse Approaches to Open Gov't Data

- Government data initiatives have taken many forms
- GovData portals are widely varied in how they help users discover and use relevant datasets
Stakeholders have seen the need for federated discovery across catalogs, especially from within major search engines including Bing, Google, Yahoo! and Yandex.
• Publishing open government data as Linked Data is not enough
• For OGD to be useful, datasets must be published using metadata, markup standards and presentation that aid discovery and use
• Publishing open government data as Linked Data is **not** enough

• For OGD to be useful, datasets must be published using metadata, markup standards and presentation that aid discovery and use
Recent work at TWC RPI demonstrates the value of applying emerging standards for uniformly describing government datasets and catalogs.
TWC's IOGDS application is an aggregated catalog of more than 1M datasets from over 192 dataset catalogs from governments at every level around the world.

See: http://logd.tw.rpi.edu
Data Catalog Vocabulary (DCAT)

W3C Editor’s Draft 11 July 2012

This version:
http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html
Latest published version:
http://www.w3.org/TR/vocab-dcat/
Latest editor’s draft:
http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html
Previous version:
none
Editors:
Fadi Maali, DERI, NUIG
John Erickson, Tethersless World Constellation (RP)
Phil Archer, W3C/ERCIM

Abstract

DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.

By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can serve as a manifest file to facilitate digital preservation.

Anticipates W3C DCAT RDF vocabulary

Demos what a comprehensive federated catalog based on DCAT and aggregation API might look like
IOGDS is a multi-year effort based on downloading, scraping or accessing APIs, converting metadata to a proto-DCAT model, and publishing via endpoint and download.

See: http://logd.tw.rpi.edu
TWC RPI has published dataset listings based on IOGDS using emerging microdata standards, esp. schema.org model endorsed by Bing, Google, Yahoo!, Yandex...
• TWC RPI's schema.org dataset extension will enable government dataset catalogs to more easily be parsed and indexed by the major search engines...

• ...which will help users find relevant datasets!

• TWC's dataset extension entered public discussion June 2012
The schema.org datasets extension enables relevant datasets to be more easily discovered by a range of stakeholders including researchers, data journalists, bloggers and developers.
“...we've reviewed the current datasets schema proposal in draft, and we are comfortable with the current state of things...

“...At this point, if the group would solidify on the dataset proposal, then Data.gov would support and use it.

---Chris Musialek
• API-based catalog federation is also possible
• ckan announced DCAT-based query/federation API
• enables OAI-PMH-like harvesting and more
• Geo-based discovery: What data is available by geo-selection?
• Provenance-based discovery: How do I get the data that someone else used? “Get the Data”
• Community/social-based discovery: Dude, check out this data! (Linked Data perfect for this...)
Other Thoughts...

- Geo-based discovery:
  What data is available by geo-selection?

DATA.gov Geo Viewer
Other Thoughts...

- Community/social-based discovery: *Dude, check out this data!*

*OPENEI.org*
Choose your own medicine...
but do expose your **metadata**
and **get** your catalogs **discovered**!