Data.gov Wiki: A Semantic Web Approach to Government Data

Li Ding, Dominic DiFranzo, Sarah Magidson, Alvaro Graves, James R. Michaelis, Xian Li, Deborah L. McGuinness, Jim Hendler

Tetherless World Constellation
Nov 2, 2009
Synergy

- Government: data is out there “as is”
- Loop: gov data and linked data
- Loop: gov data and web developers
- Loop: gov data and end users
Government Data on the Web

Welcome to Data.gov

The purpose of Data.gov is to increase public access to high value, machine readable data sets generated by the Executive Branch of the Federal Government. Although the initial launch of Data.gov provides a limited portion of the rich variety of Federal datasets presently available, we invite you to actively participate in shaping the future of Data.gov by suggesting additional datasets and site enhancements to provide seamless access and use of your Federal data. Visit today with us, but come back often, with your help, Data.gov will continue to grow and change in the weeks, months, and years ahead.

How to use Data.gov

Data.gov includes searchable data catalogs providing access to data in three ways, through the "raw" data catalog, the tool catalog and the geodata catalog. Please note that by accessing data and tools offered on Data.gov, you agree to the Data Policy, which should be read before accessing any dataset or tool. If there are additional datasets that you would like to see included on this site, please click here. For more information on how to use Data.gov, view our tutorial.
Objectives

• Investigate the role of semantic web in producing, processing and utilizing government datasets
  – To enrich the value of data via normalizing, linking and information-extraction
  – To realize the value of data via applications, esp. visualization
  – To support web developers via machine friendly data access and web services
Semantic Web Architecture for Government Data

Data Processors
(Web Services & Analyzers)

XSLT Service
Diff Service
Link Annotator
RSS Generator
...

SPARQL Web Service

RDF/XML

GOV data
(RDF)

Linked Data

CSV
XSL
...

Sem Wiki
Tabulator
Google Viz
MIT Exhibit
RSS 1.0
tagCloud

Li Ding, Dominic DiFranzo, Sarah Magidson, and Jim Hendler · Tetherless World Constellation · Rensselaer Polytechnic Institute · Aug 7 2009 · http://data-gov.tw.rpi.edu/
The Landscape
# The catalog data

<table>
<thead>
<tr>
<th>Patent Application Bibliographic Data (2009)</th>
<th>92/agency</th>
<th>Department of Commerce</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>92/agency data series page</td>
<td><a href="http://www.uspto.gov/web/menu/patdata.html">http://www.uspto.gov/web/menu/patdata.html</a></td>
</tr>
<tr>
<td></td>
<td>92/agency program page</td>
<td>Patent Application Bibliographic Data</td>
</tr>
<tr>
<td></td>
<td>92/applicable agency</td>
<td>Business Enterprise</td>
</tr>
<tr>
<td></td>
<td>information quality guideline</td>
<td>USPTO Patent Application Bibliographic Data</td>
</tr>
<tr>
<td></td>
<td>designation</td>
<td><a href="http://www.uspto.gov/web/patents/howtopat.htm">http://www.uspto.gov/web/patents/howtopat.htm</a></td>
</tr>
<tr>
<td></td>
<td>92/category</td>
<td><a href="http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/">http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/</a></td>
</tr>
<tr>
<td></td>
<td>92/citation</td>
<td>Raw Data Catalog</td>
</tr>
<tr>
<td></td>
<td>92/collection mode</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>92/data collection instrument</td>
<td>15-Mar-2001</td>
</tr>
<tr>
<td></td>
<td>92/data dictionary variable list</td>
<td>Thursdays</td>
</tr>
<tr>
<td></td>
<td>92/data gov data category type</td>
<td>Contains the bibliographic text (i.e., front page) of each</td>
</tr>
<tr>
<td></td>
<td>92/data quality certification</td>
<td>images/drawings. This file is a subset of the Patent App</td>
</tr>
<tr>
<td></td>
<td>92/date released</td>
<td>International Common Element (ICE) Document Type D</td>
</tr>
<tr>
<td></td>
<td>92/description</td>
<td>where &quot;yyyyymmdd&quot; is a Thursday publication</td>
</tr>
<tr>
<td></td>
<td>92/frequency</td>
<td>each week [where &quot;yyyyymmdd&quot; is a Thursday publication]</td>
</tr>
<tr>
<td></td>
<td>92/geographic coverage</td>
<td>are (3) files: ipabyyyyyymmd.xml (Bibliographic information)</td>
</tr>
<tr>
<td></td>
<td>92/regularity</td>
<td>For more information about these files: &lt;a href=&quot;<a href="http://w">http://w</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>e.gov/office/myo/ice/ndoe/online.html&quot;</td>
</tr>
</tbody>
</table>

---
Data-gov Cloud (Aug 2009)

Population

Population
(#90) 2005-2007 ACS PUMS Housing
(#91) 2005-2007 ACS PUMS Population

Energy and Utilities

Budget

Budget
(#401) Budget Authority and offsetting receipts 1976-2014
(#402) Outlays and offsetting receipts 1962-2014
(#403) Governmental Receipts 1962-2014

Residential Energy Consumption Survey
(#10) 1976-2014

Governmental Receipts
(#402) 1962-2014
Outlays and offsetting receipts

Worldwide M1+ Earthquakes past 7 days
(#34)

2006 Toxics Release Inventory
(#249)

2005 Toxics Release Inventory
(#191)

2007 Toxics Release Inventory
(#397)

CASTNET Visibility
(#9)

CASTNET Ozone
(#8)

CASTNET sites
(@10001)

Worldwide M1+ Earthquakes past 7 days
(#34)

2006 Toxics Release Inventory
(#249)

2005 Toxics Release Inventory
(#191)

2007 Toxics Release Inventory
(#397)

CASTNET Visibility
(#9)

CASTNET Ozone
(#8)

CASTNET sites
(@10001)

Worldwide M1+ Earthquakes past 7 days
(#34)

2006 Toxics Release Inventory
(#249)

2005 Toxics Release Inventory
(#191)

2007 Toxics Release Inventory
(#397)

CASTNET Visibility
(#9)

CASTNET Ozone
(#8)

CASTNET sites
(@10001)
More statistics
Demos
Data.gov + epa.gov
Gov Data + Corporate Data + User Data

Annotated Timeline for USPS Money and Historical Events

A. Outlays for Postal Service Fund 1.44 b
B. Net Income 1.04 b
C. Budget Authority for Postal Service Fund 2.24 b

E. National Postal Museum opens
   New museum showcases US postal history. It is funded by the USPS, the Smithsonian Institute, and private donations.
   1993-7-30

D. Government Performance and Results Act
   This act requires government agencies to periodically release a 5-year plan on goals and how they plan to achieve them.
   1993-1-5

F. John E. Potter becomes postmaster
   He is the 72nd Postmaster General of the US.
   2001-6-1

G. 2002 Transformation Plan released
   The USPS releases a report on major structural changes and reforms it can make to meet the challenges of the 21st century, as well as recommendations to Congress on how USPS can be helped.
   2002-4-1

H. Union rise in net income:
   2003-4-23

1. 1d 2d 3m 3m 5m Max
2. Zoom: 1d 2d 3m 5m Max
3. Jan 05, 1993
4. Outlays for Postal Service Fund 1.44 b
5. Net Income 1.04 b
6. Budget Authority for Postal Service Fund 2.24 b
7. Union rise in net income:
   2003-4-23

Net Income = Total income - total expenses
Computing Difference of Revisions


[new] Savings Bond Issues, Redemptions, and Maturities by Series


--- > xls_access_point . [values] [http://www.treasurydirect.gov/govt/reports/pd/pd_redemptionissuesbyseries.xls]

--- > agency . [values] [Department of the Treasury]

--- > statistical_characteristics . [values] [Not Relevant]

--- > data_quality_certification . [values] [Yes]

--- > applicable_agency_information_quality_guideline_designation . [values] [Treasury]

--- > statistical_methodology . [values] [Not Relevant]

--- > questionnaire_design . [values] [Not Relevant]

--- > privacy_and_confidentiality . [values] [Not Relevant]

--- > title . [values] [Savings Bond Issues, Redemptions, and Maturities by Series]
More demos?

Technical Issues
Issues in Data.gov

- **Duplicated Datasets** - Some datasets are part of another dataset
  - Dataset 140 (2005 Toxics Release Inventory data for the state of California (EPA)) is a subset of Dataset 191.

- **Formatting Issues** - The format of some datasets is not friendly to machine processing.
  - Dataset 37 (Lower Colorado River Daily Average Water Elevations and Releases (US Bureau of Reclamation)).
  - Dataset 335 (National Longitudinal Surveys (US Bureau of Labor Statistics)) tells you how to order data from the government.

- **Access Point Issues** - The access points are interactive webpage which is not friendly for machine access.
  - Dataset 330 (Local Area Unemployment Statistics (US Bureau of Labor Statistics))
Linking Data

1. link similar datasets by reusing property namespace
2. link to rdfs:label (via rdfs:subPropertyOf) using semantic wiki
3. link to DBpedia (via owl:sameAs) using wikipedia widget
4. link instances (via common <property, literal-value> pair)
5. link government data with web data (via time and location)
6. link revisions of government data (via knowledge provenance)
Semantic mapping: AI + CI

need manual disambiguation!
RDF => SPARQL => Web

- We use SPARQL to bridge Web developers and Semantic Web data.
- A triple store is used to support handling multi-million triple RDF datasets.
Conclusion

- semantic web enabled portal for linked government data
- 5 billion triples from data.gov
- hosts apps, demos & services
- provide education services
- integrates web users’ contributions