Peta vs. Meta: Rethinking Data Interoperability on the World Wide Web

Jim Hendler
Director, Rensselaer Institute for Data Exploration and Applications
Rensselaer Polytechnic Institute, USA
http://www.cs.rpi.edu/~hendler
@jahendler (twitter)
Geekopedia: Data exploration helps a data consumer focus an information search on the pertinent aspect of relevant data before true analysis can be achieved. In large data sets, data is not gathered or controlled in a focused manner. Even in smaller data sets, it is also true that data gathered are not in a very rigid and specific technique can result in a disorganized manner and a myriad of subsets each...
Tetherless World Constellation

Government Data Sharing

International OGD Catalog Search, searching 440,386 datasets from 116 catalogs in 16 languages representing 38 countries and international organizations.

Results 1 to 200 of 440386

Dataset Title | Agency | Categories | Catalog | Country
--- | --- | --- | --- | ---
Elektronischer Wasserstraßen-Informationsservice - [Karte Nordwest] | Niedersächsischer Landesbetrieb für Wasserwirtschaft Kosten- und | Wasser | Portalu.de (Germany) | Germany

International OGD Catalog Search, searching 1,022,787 datasets from 192 catalogs in 24 languages representing 43 countries and international organizations.
Semantic Web and Linked Data (UK)

<table>
<thead>
<tr>
<th>Potholes in our county council</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date reported</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>15/01/11 High Street</td>
</tr>
<tr>
<td>16/01/11 Tree Lane</td>
</tr>
<tr>
<td>16/01/11 High Street</td>
</tr>
<tr>
<td>18/01/11 B3068</td>
</tr>
<tr>
<td>19/01/11 B3068</td>
</tr>
<tr>
<td>22/01/11 B2013</td>
</tr>
<tr>
<td>23/01/11 High Street</td>
</tr>
<tr>
<td>24/01/11 High Street</td>
</tr>
<tr>
<td>24/01/11 London Road</td>
</tr>
<tr>
<td>26/01/11 Birmingham Road</td>
</tr>
<tr>
<td>27/01/11 High Street</td>
</tr>
<tr>
<td>28/01/11 High Street</td>
</tr>
</tbody>
</table>

County Council

<table>
<thead>
<tr>
<th>Street</th>
<th>Post code</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Street</td>
<td>BM1 3</td>
</tr>
<tr>
<td>Tree Lane</td>
<td>BM1 2</td>
</tr>
<tr>
<td>B3068</td>
<td>BM1 1</td>
</tr>
<tr>
<td>London Road</td>
<td>BM1 2</td>
</tr>
<tr>
<td>Birmingham Road</td>
<td>BM1 3</td>
</tr>
</tbody>
</table>

Royal Mail

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BM1 1AE</td>
<td>291980</td>
</tr>
<tr>
<td>BM1 1AT</td>
<td>291778</td>
</tr>
<tr>
<td>BM1 1BA</td>
<td>291725</td>
</tr>
<tr>
<td>BM1 1BB</td>
<td>291786</td>
</tr>
<tr>
<td>BM1 1BD</td>
<td>291763</td>
</tr>
</tbody>
</table>

Ordnance Survey
Government Data is currently over ½ the cloud in size (~17B triples), 10s of thousands of links to other data (within and without)

http://linkeddata.org/
Creating Data Mashups Requires Semantics

Tetherless World Constellation

More than 50 of these at http://logd.tw.rpi.edu
Head to head comparisons shows that burglaries in Avon and Somerset (UK) far exceed those in Los Angeles, California (one of the highest crime areas in the US)
The problem is (likely) semantics

Do the terms mean the same? Are they collected in the same way? Are they processed differently? …
• International data sharing
  – W3C Govt Linked Data Working Group
  – Need for vocabularies within govt sectors
    • Esp for cross-language use
      – How can we compare health (or legal, or social, or ....) data between countries like US, UK, India, Kenya (English) with Norway, China, France, etc.
      – How can we link local govts (in traditional languages, local dialects, etc) w/national data

• Modern Metadata design is crucial to govt data sharing
  – Needed for search and federation in large data sharing efforts
Traditionally metadata tries to be comprehensive
- Example: ISO 19115 (GIS standard)
  - >400 elements
  - 14 “packages”
  - Dozens of UML models (not all consistent with each other)
• Big Data on the web
  – is moving away from traditional relational models (cf. NoSQL)
  – Moving towards third party application and extension (cf. Json)

  – Focus on interoperability and exchange with “lightweight” semantics
  • Using ideas from the Semantic Web
    – Search: Schema.org
    – Social Networking: OGP
Tetherless World Constellation

Schema.org
### Dataset extension to schema.org - April, 2013

#### Tetherless World Constellation

<table>
<thead>
<tr>
<th>Datasets extension</th>
<th>DCAT</th>
<th>ADMS</th>
<th>VoID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ds:DataCatalog</td>
<td>dcat:Catalog</td>
<td>adms:SemanticAssetRepository</td>
<td></td>
</tr>
<tr>
<td>ds:DataDownload</td>
<td>dcat:Download</td>
<td>adms:SemanticAssetDistribution</td>
<td></td>
</tr>
<tr>
<td>ds:Dataset</td>
<td>dcat:Dataset</td>
<td>dcterms:PartOf</td>
<td></td>
</tr>
<tr>
<td>ds:catalog</td>
<td>dcterms:hasPart</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ds:dataset</td>
<td>dcat:dataset</td>
<td>dcterms:license</td>
<td></td>
</tr>
<tr>
<td>ds:spatial</td>
<td>dcterms:spatial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:about</td>
<td>dcat:theme</td>
<td>dcterms:subject</td>
<td></td>
</tr>
<tr>
<td>sdo:contentSize</td>
<td>dcat:size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:contentURL</td>
<td>dcat:accessURL</td>
<td>adms:accessURL</td>
<td></td>
</tr>
<tr>
<td>sdo:copyrightHolder</td>
<td>dcterms:license</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:Country</td>
<td>dcterms:created</td>
<td>dcterms:modified</td>
<td></td>
</tr>
<tr>
<td>sdo:dateModified</td>
<td>dcterms:modified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:datePublished</td>
<td>dcterms:issued</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:description</td>
<td>dcterms:description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:encodingFormat</td>
<td>dcterms:format</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:inLanguage</td>
<td>dcterms:language</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:name</td>
<td>dcterms:title</td>
<td>rdfs:label</td>
<td></td>
</tr>
<tr>
<td>sdo:Organization</td>
<td>foaf:Organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:Person</td>
<td>foaf:Person</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:publisher</td>
<td>dcterms:_publisher</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:Thing</td>
<td>skos:Concept</td>
<td>(recommends but does not require skos:Concept)</td>
<td></td>
</tr>
<tr>
<td>sdo:url</td>
<td>foaf:homepage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sdo:version</td>
<td>dcat:CatalogRecord</td>
<td></td>
<td>dcat:datasetDictionary</td>
</tr>
</tbody>
</table>

**Schema.org/Dataset** – add this to your pages!
**SF Shoreline**
San Francisco mainland shoreline and in the south, the county line.

*Country: United States*

*Publisher: Department of Public Works*

Human-readable database description (HTML)
type: http://schema.org/dataset

name: SF Shoreline


description: San Francisco mainland shoreline and in the south, the county line.

spatial: Item 9

publisher: Item 10

Item 9

type: http://schema.org/country

name: United States

Item 10

type: http://schema.org/organization

Department of Public Works
Tetherless World Constellation

Schema.org/Dataset
google.com/publicdata (early days)

World Development Indicators
World Bank
This dataset contains the World Development Indicators (WDI).

IFs Forecast - Version 6.63
Frederick S. Pardee Center for International Futures
International Futures (IFs) Forecasts in cooperation with the Strategic Foresight Project of the Atlantic Council and the US National Intelligence Council

Human Development Indicators
Human Development Report 2013, United Nations Development Programme
The data used for calculating the Human Development Index (HDI) and other composite indices featured in the Human Development Report are provided by a ...
The White House
Office of the Press Secretary

For Immediate Release
May 09, 2013

Executive Order -- Making Open and Machine Readable the New Default for Government Information

EXECUTIVE ORDER

MAKING OPEN AND MACHINE READABLE THE NEW DEFAULT FOR GOVERNMENT INFORMATION

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:
[{
  "title": "Data Catalog",
  "description": "Version 1.0",
  "keyword": "catalog",
  "modified": "2013-05-09 06:00:00",
  "publisher": "US Department of X",
  "person": "Contact Person",
  "mbox": "contact.person@agency.gov",
  "identifier": "1",
  "accessLevel": "public",
  "distribution": [
    {
      "accessURL": "http://agency.gov/data.json",
      "format": "json"
    }
  ]
},
{
  "title": "Public Elementary/Secondary Listing",
  "description": "The purpose of the CCD nonfiscal surveys is to provide a listing of all schools and agencies providing education, schools",
  "modified": "2011-11-19 00:00:00",
  "publisher": "US Department of Education",
  "person": "Open Data Initiative",
  "mbox": "opendata@ed.gov",
  "identifier": "ykv5-fn9t",
  "accessLevel": "public",
  "distribution": [
    {
      "accessURL": "https://explore.data.gov/views/ykv5-fn9t/rows.csv?accessType=DOWNLOAD",
      "format": "csv",
      "size": "20mb"
    },
    {
      "accessURL": "https://explore.data.gov/views/ykv5-fn9t/rows.json?accessType=DOWNLOAD",
      "format": "json"
    },
    {
      "accessURL": "https://explore.data.gov/views/ykv5-fn9t/rows.xml?accessType=DOWNLOAD",
      "format": "xml"
    }
  ],
  "webService": "http://explore.data.gov/api/views/ykv5-fn9t/rows.json",
  "license": "Public Domain",
  "spatial": "US",
  "temporal": "2009-09-01 00:00:00,2010-05-31 00:00:00",
  "issued": "",
  "frequency": "one-time",
  "language": "English",
  "granularity": "
  "dataQuality": "true",
  "theme": "education",
  "references": "http://nces.ed.gov/ccd/data/text/peu091alay.txt",
  "landingPage": "http://ed.gov/developer",
  "feed": "",
  "systemOfRecords": "http://nces.ed.gov/ccd/
}
]
Office Locations

- **Title:** Office Locations
- **Description:** A list of the agency's office locations and contact information.
- **Documentation URL:** http://www.agency.gov/data/information/locations
- **Download URL:** http://www.agency.gov/data/raw/locations.zip
- **Format:** csv
- **Tags:** keyword1
- **Last Update:** 1/1/2013
- **Publisher:** Agency
- **Contact Name:** John, Smith
- **Contact Email:** john.smith@agency.gov
- **Unique Identifier:** 1
- **Public:** true
- **Endpoint:** http://www.agency.gov/data/raw/locations.json
- **License:** public domain
- **Spatial:** United States
- **Temporal:** today
- **Release Date:** 7/9/2012
- **Frequency:** 6 months
- **Language:** English
- **Granularity:** Address
- **Category:** Energy
- **Related Documents:** http://www.agency.gov/data/information/locations/document.doc
- **Distribution:**
  - **Size:** 44KB
- **Homepage URL:**
- **RSS Feed:**
- **Data Quality:** True
EU moving in right direction
It’s not enough just to describe the data elements…

Person

Date

Smith, James: June 4
Jones, Fred: May 17
O’Connell, Frank: April 3
Chang, Wu: February 21
Hoffman, Bernd: December 9
Describing a dataset ... requires a context

Tetherless World Constellation

1976 Dates of Birth

Person: Smith, Jones, O’Connell, Chang, Hoffman
Date: June 4, May 17, April 3, February 21, December 9
Describing a dataset ... requires a context
How do we capture more of this information?

Tetherless World Constellation

1976 Cancer Mortality dates

Person

Smith  James  June 4
Jones  Fred  May 17
O’Connell  Frank  April 3
Chang  Wu  February 21
Hoffman  Bernd  December 9
Linked Semantic Web “metadata” documents can be used to link very large databases in distributed data systems. This leads to orders of magnitude reduction in information flow for large-scale distributed data problems.
Conclusions

• Open data is becoming “Broad Data”
  – World Wide Web trend towards more and more varied data
    • In many domains
      – E-commerce, Open Govt, many more (cf. Health/Medical care)

• Broad data requires thinking outside the “Database” box
  – DIVE: discover, integrate, visualize, explain

• Broad data requires
  1. Modern, Web-oriented metadata
  2. LINKING the metadata, not the data