- Watch the web site for schedule (especially wrt Nov 20)!

- Read the paper on „web science“ as indicated on http://west.uni-koblenz.de/teaching/ws1213/webscience
  it will be considered part of the curriculum and will be asked for in the oral exam
Structured Data on the Web

Introduction to Web Science

Steffen Staab

Slides from Jim Hendler & Semantic Web Company
Agenda

- HTML Meta
- Microdata
- Microformats
- RDF
  - RDF-a

- Applications
  - Schema.org – done already
  - Wikidata - seminar
  - Open data (especially Open government data)
Open Government Data is a worldwide movement to open data (& information) of the government / public administration* - that is NOT personal (individual related) – in human- AND maschine-readable open (non proprietary) formats for re-use (by civil society, economy, media, academia AND politics & public administration itsself)!

OPEN stands for lowering the barriers to ensure as broad as possible re-use (for everybody)!

There is a new paradigm in publishing Open Government Data = look, take and play!

* ..... data and information produced or commissioned by government or government controlled entities
1. Completeness
All public data are made available. Public data are data that are not subject to valid privacy, security or privilege limitations, as governed by other statutes.

2. Primacy
Data are published as collected at the source, with the finest possible level of granularity, not in aggregate or modified forms.

3. Timeliness
Data are made available as quickly as necessary to preserve the value of the data.

4. Ease of Physical and Electronic Access
Data are available to the widest range of users for the widest range of purposes.

5. Machine readability
Data are reasonably structured to allow automated processing of it.

6. Non-discrimination
Data are available to anyone, with no requirement of registration.

7. Use of Commonly Owned Standards (e.g. open formats)
Data are available in a format over which no entity has exclusive control.

8. Licensing
Data are not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed as governed by other statutes.

Finally, compliance must be reviewable.

PLUS: Sunlight Foundation, August 2010: 9) Permanence & 10) Usage Costs
5 Stars for Open Data by Tim Berners Lee

- ⭐️ make your stuff available on the web (whatever format)
- ⭐️⭐️ make it available as structured data (e.g. excel instead of image scan of a table)
- ⭐️⭐️⭐️ non-proprietary format (e.g. csv instead of excel)
- ⭐️⭐️⭐️⭐️ use URLs to identify things, so that people can point at your stuff
- ⭐️⭐️⭐️⭐️⭐️ link your data to other people’s data to provide context

From Open Data to Linked Open Data

Publish Data!
• Organise Data!
• License Data!
• Raw Data now!

Use Web-Technologies
• Provide an API!

Use Linked Data!
• Give things an URI!
• Use RDF for Publishing!
• Link your Data to other Data (as well as the data models)!
• Provide a Standard-API

• The web is an Ecosystem
• Networked Data creates Network Effects
• Lowers Costs of Data Integration

Data have swept into every industry and business function and are now an important factor of production, alongside labor and capital.

Harnessing big data in the public sector has enormous potential, too. If US health care were to use big data creatively and effectively to drive efficiency and quality, the sector could create more than $300 billion in value every year. Two-thirds of that would be in the form of reducing US health care expenditure by about 8 percent.

In the developed economies of Europe, government administrators could save more than €100 billion ($149 billion) in operational efficiency improvements alone by using big data, not including using big data to reduce fraud and errors and boost the collection of tax revenues.

And users of services enabled by personal location data could capture $600 billion in consumer surplus.

EXAMPLES FROM USA / RENSSELAER POLYTECHNIC INSTITUTE
Government Data on the Web

Welcome to Data.gov

The purpose of Data.gov is to increase public access to high value, machine readable datasets 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 geospatial catalog. Please note that by accessing data tools offered on Data.gov, you agree to the Data Policy, which should read before accessing any dataset or tool. If there are additional datasets that you would like to see included on this site, please let us know. For more information on how to use Data.gov, view our tutorial.

staab@uni-koblenz.de
"Openness will strengthen our democracy and promote efficiency and effectiveness in Government."
--- President Obama

2009

January 1, 2009
WeST
staab@uni-koblenz.de

January 19, 2010

May 21, 2009

May 21, 2010

June 30, 2009

December 8, 2009

Putting Govt Data online-
Data.gov.uk beta

data.gov online
57 Data Sets

“Open Government Directive” released
~2000 Data Sets

data.gov relaunch with semantic web featured
>305,000 Data Sets

~6000 Data Set

2010 …
Important to the citizens: eg. Education
Moving data.gov to linked data (UK)

- Built around “linked data” from the start
- Authorization for this from the Prime Minister
• Third parties (like RPI) translate the government datasets into linked data formats

• US Data.gov hosts 6.4B RDF triples 5/21/2010
  • Semantic Web community hosted
  • http://data.gov/semantic
Linked data lets us create “Data” Mashups

More than 50 of these at http://logd.tw.rpi.edu
(and lots more at data.gov.uk)
Data.gov + epa.gov

Castnet Ozone Map

Ozone Stations

Land Use
- 27 Agric
- 1 CoastalMarsh
- 3 Desert
- 40 Forest
- 1 ForestArid
- 2 ForestMarsh

Terrain
- 37 Rolling
- 4 MountainTop
- 1 FlatWater
- 14 Flat
- 27 Complex
- 1 Coastal

Agency
- 59 EPA
- 25 NPS

Wind Cave NP from NPS
53ppb (Average Ozone level)
Land Use: Prairie
See Data
More Information

Ozone order
- 30
- 36
- 42
- 48
- 54
- 60

Map data ©2009 Tele Atlas, INEGI - Terms of Use
A Typical Mashup: CASTNET

1. Convert raw dataset into linkable RDF
2. Query multiple RDF dataset via SPARQL end point
3. Drill down for details
4. Surf to EPA applications

Data Mashup
Visualization Mashup
Web Application Mashup

Created by Dominic DiFranzo, PhD student at RPI, http://www.data.gov/semantic/Castnet/html/exhibit
Adding some Web magic
Linking GDP of the US and China

Federal Reserve Statistical Release
H.10
Foreign Exchange Rates (Weekly)

GDP of the US (Billion Dollar)

GDP of China (Billion Chinese Yuan)

Linking GDP of the US and China

This mashup was built in less than 4 hours – including conversion of data, web interface, and visualization!

Datasets: acres burned, and agency budgets
Dbpedia: wikipedia descriptions of major US fires
Integrate with Social media

Linking Open Government Data

Team Endpoint

Description: This map shows some locations that receive funding from the New York State Government. Click on a pin to learn more about that location and cast your votes via Twitter.

Cast Your Votes on New York State Projects

238 items

40 results out of 238 cannot be plotted.

Search Projects

City
1. Albertson
2. Baldwin
3. Bellmore
4. Bethpage
5. East Hills

Agency
16. Aging
6. ASAS
6. C&F
53. CPS
1. DCJS

East Meadow LAX

Description: To help defray operating expenses and provide programs for the youth.

Address: 2510 Second Avenue
East Meadow, NY 11554

Funding: $1750

Senator: Hannon

For: 2

Against: 2

Steffen Staab
staab@uni-koblenz.de
Combining data from different data sharing sites

Comparing US (USAID) and UK (DFID) Global Foreign Aid

Aid Figures

- Aid to Indonesia (in USD)
  - USAID
  - DFID

USAID Categories (in USD)
- Economic/Social Assistance
- Development Assistance
- Child Survival and Health
- Other USAID Assistance

DFID Categories (in USD)
- Education
- Health
- Social Services
- Government and Civil Society
- Economic
- Research
- Humanitarian Assistance
- Non Sector Allocable
RPI workflow enhances raw RDF w/ useful URIs

- **Convert**
- **Access**
- **Enhance**
- **derive**
- **create**
- **Version**
- **revision**
- **SemDiff**
- **derive**

Steffen Staab
staab@uni-koblenz.de
http://logd.tw.rpi.edu
• demos, tutorials, RDF-ized datasets,

Featured LOGD Demos

Contributor(s):

Quick Links
• Visit our new International Dataset Catalog
• Semantic Web Challenge 2010 - TWC wins 2nd prize at SWC 2010!!!
• Demos - Examples of consuming open government data
• Tutorials - Learn Semantic Web technologies for LOGD
• Feedback - Send us comments
• Issue tracker - Report a bug
• LOGD Open Source on Google Code
• LOGD SPARQL Endpoint (Alpha)

TWC LOGD Datasets Stats (as of 2011-11-04)
• TWC LOGD is hosting 9,951,771,397 RDF triples
• We have created 1,888 RDFized datasets originating from 119 sources.
• We have enhanced 1,686 RDFized datasets using 9,499 links to other LOD datasets.
• For more statistics, see TWC LOGD: Data Description and Statistics.

US Global Foreign Aid

LOGD Portal Updates (aggregated RSS Feeds from several sources)

LOGD Videos (source: LOGD site)
• Analyzing Political Topics Using Open Government Data
  by dingo - Sep 27, 2010
TWC LOGD Datasets Stats (as of 2011-11-04)

- TWC LOGD is hosting 9,951,771,397 RDF triples
- We have created 1,888 RDFized datasets originating from 119 sources.
- We have enhanced 1,686 RDFized datasets using 9,499 links to other LOD datasets.
- For more statistics, see TWC LOGD: Data Description and Statistics.
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/
URI design

- URI design is crucial to govt data sharing
  - esp. within govts
  - Whether your goal is linked data or not
- UK Government has designed and made great use of standard URI practices in their linked data
  - US exploring URI design schemes
    - Join the community at Semantic.data.gov and participate!
A well-known benefit of using Linked Data principles when publishing government data is the ease with which the resulting data may be integrated and linked with diverse other datasets. Realizing such benefits ultimately hinges on unambiguously identifying common entities within these datasets. From a practical standpoint this means identifying which literals represent instances of entities --- US states, US government agencies, toxic chemicals, crops, etc --- and associating canonical URIs with those occurrences across datasets. To facilitate interlinking, it is then helpful to relate these literals and our newly-minted canonical identifiers to other identifiers in common usage --- DBPedia data pages, information pages used in various industries, etc --- to further facilitate interlinking with other datasets.

The TWC LOGD Instance Hub Project seeks to demonstrate the utility of this approach for entities occurring across diverse government datasets. First, we have identified several categories of entities from Data.gov datasets. We have designed canonical URIs based on a draft set of principles that may be applied across a wide range of government data sources (agencies, etc). And we have provided links to synonymous identifiers for these entities in common use in other systems to facilitate interlinking.

TWC has established a draft set of URI Design Principles that government agencies may choose to adopt as they establish authoritative, canonical URIs for entities of special concern within their datasets. The TWC LOGD Instance Hub demonstrates how to associate these canonical URIs with instances of entities that occur as literals or as URIs from other naming systems.

The following table provides links to instance data for a number of categories derived from US Government data via Data.gov. Click on a category to browse these instances or download the RDF data (in Turtle).

<table>
<thead>
<tr>
<th>Instance Hub Category</th>
<th>RDF Dump</th>
</tr>
</thead>
<tbody>
<tr>
<td>US States</td>
<td>.ttl</td>
</tr>
<tr>
<td>US Agencies</td>
<td>.ttl</td>
</tr>
<tr>
<td>US Fiscal Years</td>
<td>.ttl</td>
</tr>
<tr>
<td>Crops</td>
<td>.ttl</td>
</tr>
<tr>
<td>Toxic Chemicals</td>
<td>.ttl</td>
</tr>
<tr>
<td>Alternative Transport Fuels, etc (Coming Soon)</td>
<td>.ttl</td>
</tr>
</tbody>
</table>
Example: US States

New York

Instance Hub: New York
http://lod.tw.rpi.edu/id/us/state/New_York

Instance Hub State Page
Instance Hub Main Page

Linked To:
- http://sw.opencyc.org/concept/Mw4rV/VNs5wpEbGdrcNSY29ycA
- http://rdf.freebase.com/ns/m/099rby
- http://rdf.freebase.com/ns/m/01mifp

Datasets Linked:
Example: US Govt Agencies

National Oceanic and Atmospheric Administration

Instance Hub: National Oceanic and Atmospheric Administration

http://lod.tw.rpi.edu/id/us/fed/agency/Department_of_Commerce/National_Oceanic_and_Atmospheric_Administration

Instance Hub US Agency Page

Instance Hub Main Page

Abbreviation: NOAA; DOC/NOAA

The National Oceanic and Atmospheric Administration (NOAA), pronounced /noʊə/ like "noah", is a scientific agency within the United States Department of Commerce focused on the conditions of the oceans and the atmosphere. NOAA warns of dangerous weather, charts seas and skies, guides the use and protection of ocean and coastal resources, and conducts research to improve understanding and stewardship of the environment. In addition to its civilian employees, NOAA research and operations are supported by 300 uniformed service members who make up the NOAA Commissioned Officer Corps. The current Under Secretary of Commerce for Oceans and Atmosphere at the Department of Commerce, and the agency’s Administrator, is Dr. Jane Lubchenco, nominated by President Barack Obama and confirmed by the United States Senate on March 19, 2009.

Homepage: http://www.noaa.gov/

owl:SameAs: http://dbpedia.org/resource/National_Oceanic_and_Atmospheric_Administration

http://rdf.freebase.com/ns/m/09fyq

Datasets Linked:

Instance Hub: soybean

http://logd.tw.rpi.edu/source/usda-gov/id/us/crops/soybean

Instance Hub US Crop Page
Instance Hub Main Page

Crop Category:      oil crops

USDA Briefing URL: http://www.ers.usda.gov/Briefing/SoybeansOilCrops/
http://dbpedia.org/resource/Soybean

owl:SameAs:
http://eunis.eea.europa.eu/species/171391
http://rdf.freebase.com/ns/m/0gyv9
http://lod.geospecies.org/identifier/52Dxq

Datasets Linked:
Metadata design

- Metadata design is crucial to govt data sharing
  - Needed for search and federation in large data sharing efforts
- International data sharing will be a crucial next step
  - W3C Govt Linked Data Working Group
  - Need for vocabularies within govt sectors
    - Esp for cross-language use
### 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

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Agency</th>
<th>Categories</th>
<th>Catalog</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elektronischer Wasserstraßen-</td>
<td>Niedersächsischer Landesbetrieb f r Wasserwirtschaft Kisten- und Naturschutz</td>
<td>Wasser</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Informationsservice - [Karte Nordwest]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hochwasserrückhaltebecken online</td>
<td>Niedersächsischer Landesbetrieb f r Wasserwirtschaft Kisten- und Naturschutz</td>
<td>Wasser</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Bergbau - Kontrollergebnisse</td>
<td>Landesamt f r Geologie und Bergwesen Sachsen-Anhalt LAGB</td>
<td>Weitere</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Lufthygienische Überwachung Schleswig Holstein</td>
<td>Lufthygienische bewachung Schleswig-Holstein</td>
<td>Luft</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Überwachung der Umweltradioaktivität (IMIS)</td>
<td>Bundesamt f r Strahlenschutz</td>
<td>Strahlung</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Hochwasserinfos für Brandenburg</td>
<td>Ministerium f r Umwelt Gesundheit und Verbraucherschutz des Landes Brandenburg</td>
<td>Wasser</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Aktuelle Pegel der Elbe in Brandenburg</td>
<td>Ministerium f r Umwelt Gesundheit und Verbraucherschutz des Landes Brandenburg</td>
<td>Wasser</td>
<td>Portalu.de (Germany)</td>
<td>Germany</td>
</tr>
</tbody>
</table>

---

**Contact:**

Steffen Staab
staab@uni-koblenz.de
There’s lots of data out there!!

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

<table>
<thead>
<tr>
<th>Dataset Title</th>
<th>Agency</th>
<th>Categories</th>
<th>Catalog</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elektronischer Wasserstraßen-Informationsservice</td>
<td>Niedersächsischer Landesbetrieb für Wasserschutz</td>
<td>Wasser</td>
<td>Portal.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>[Karte Nordwest]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bergbau - Kontrollegebnisse</td>
<td>Geologie und Bergwesen Sachsen-Anhalt LAGB</td>
<td>Weitere</td>
<td>Portal.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Lufthygienische Überwachung Schleswig-Holstein</td>
<td>Lufthygienische Überwachung Schleswig-Holstein</td>
<td>Luft</td>
<td>Portal.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Überwachung der Umweltaktivität (UMS)</td>
<td>Bundesamt für Strahlenschutz</td>
<td>Strahlung</td>
<td>Portal.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Hochwasserinfo für Brandenburg</td>
<td>Ministerium für Umweltgesundheit und Verbraucherschutz des Landes Brandenburg</td>
<td>Wasser</td>
<td>Portal.de (Germany)</td>
<td>Germany</td>
</tr>
<tr>
<td>Aktuelle Pegel der Elbe in Brandenburg</td>
<td>Ministerium für Umweltgesundheit und Verbraucherschutz des Landes Brandenburg</td>
<td>Wasser</td>
<td>Portal.de (Germany)</td>
<td>Germany</td>
</tr>
</tbody>
</table>
Searching for data

- Faceted browser with
  - Keyword search
  - Catalogs
  - Countries
  - Agencies
  - Categories
  - (in any order)
Details and download...

- Our metadata model is described [here](http://logd.tw.rpi.edu/demo/international_dataset_catalog_search).
- An overview of our procedure for getting and publishing metadata is described [here](http://logd.tw.rpi.edu/demo/international_dataset_catalog_search).
- **NEW!** XML/RDF "dump" of IOGDS dataset catalog downloadable [here](http://logd.tw.rpi.edu/demo/international_dataset_catalog_search) (6 GB).
- **NEW!** Turtle-formatted "dump" of IOGDS dataset catalog downloadable [here](http://logd.tw.rpi.edu/demo/international_dataset_catalog_search) (1 GB).
Research remains to be done...
(it ain’t all hackathons and contests)

- Trust
  - Government data is controversial, and potentially biased
    - How do we confirm or dispute?

- Combination
  - When we combine data we need to keep the provenance of information (see trust)
    - How can we show and use?

- Scaling
  - Our project has already converted 9.9B triples from only >2,000 of the 440,000 government databases we can identify (116 catalogs, 38 countries, 16 languages)

- Versioning and updating
- Archiving
- Visualization
- ...
Exploring new visualizations

Board Members Network
Colored clusters represent boards of various US companies. Mouse-over nodes to see board member information. Click nodes to go to member’s LittleSis page. Multi-colored nodes denote board members that are on multiple boards.

Data from http://littlesis.org
Open Govt data is a critical resource
- Government data released as RDF (UK)
- Government data converted to RDF (US)
- Government data that can be found in many forms and used or converted (WWW)

Government transparency comes through in the “mashing up” of data from many datasets
- Key to linked data

An amazing opportunity for technologists (public and private) to play in an important area of the public good
- Innovation needed!
EPILOG
You've never seen data presented like this. With the drama and urgency of a spectacular slideshow, Hans Rosling debunks myths about the so-called "developing world."

In Hans Rosling's hands, data sings. Global trends in health and economics come to vivid life. And the big picture of global development—with some surprisingly good news—narrows into sharp focus. Full talk →
Now supported by 7% of web pages crawled by google

About 2% by those crawled by Yandex

Richer ontology just included (Good Relations)

No namespaces

„we will support what is popular“ Peter Mika, Yahoo
Winner of the Semantic Web Challenge 2012