Linked Open Data
A Way to Extract Knowledge from Global Datastores

Bebo White
SLAC National Accelerator Laboratory

HKU Expert Address
18 September 2014
“Developments in science and information processing have changed the meaning of the verb, ‘to know.’ It used to mean ‘having information stored in one’s memory.’ It now means the process of having access to information and knowing how to use it.”

---Herbert Simon
How can we build *knowledge* from all the facts/data on the Web?

Knowledge means data with context
Data

Information

Knowledge

Wisdom
Perceptions of Web Content

- The Web is generally thought of being composed of pages, documents
- We have been able to insert some data
  - Images `<img src="....">`
  - Multimedia
- Web 2.0 mashups provided a new way of thinking about a “Web of Data” but it was awkward to obtain
  - APIs
  - “Screen-scraping”
The Web of Documents

- Analogy
  - A global filesystem
- Designed for
  - Human consumption
- Primary objects
  - Documents (or sub-parts of)
- Links between
  - Documents (or sub-parts of)
- Degree of structure in objects
  - Fairly low
- Semantics of content and links
  - Implicit
The Web of Documents: Issues

• Simplicity
  • Loosely structured data, untyped links, disconnected data

• Integration
  • Show me all the publications from HKU PhD students in Computer Science

• Querying
  • Which papers have I written with colleagues outside the US?
The Web of Linked Documents
“Data Silos” on the Web
“Data Silos” on the Web
How About Open Data?

(1/2)

• **Interoperability** to ensure broad and easy use

• Human AND machine readable, i.e., data + metadata

• In common **open formats** using **open standards**

• Smooth and **cost efficient data integration**, i.e., reuse

• Can **generate effects** - local, regional, national, global
How About Open Data?

(2/2)

• Anyone can publish it
  • Individuals
  • Companies/Institutions
  • Governments

• Who can use it?
  • Politicians and decision makers
  • Public administration and project developers
  • Citizens (citizen analysts)
  • Economy and Industry
  • (Data) journalists, media, and publishers
  • Academia and Science
The home of the U.S. Government’s open data
Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED
SEARCH OVER 108,606 DATASETS

Monthly House Price Indexes

BROWSE TOPICS

Agriculture  Climate  Education  Energy  Finance  Geospatial

Global Development  Health  Jobs & Skills  Public Safety  Science & Research  Weather

MORE TOPICS
A World Wide Network of Data Silos
5 Stars for Open Data
(Tim B-L)

- Make content available on the Web (whatever format) - 1 ★
- Make content available as structured data - 2 ★
- Use a non-proprietary format - 3 ★
- Use URLs to identify things - 4 ★
- Link data to other data to provide context - 5 ★
The Web of Linked Data

- Analogy
  - A global database
- Designed for
  - Machines first, humans later
- Primary objects
  - Things (or descriptions of things)
- Links between
  - Things
- Degree of structure in (descriptions of ) things
  - High
- Semantics of content and links
  - Explicit
The Web of Linked Data

Don’t just link the *documents*, link the *things*
Imagine...

- A “Web” where
  - Documents are available for download on the Internet
  - But there would be no hyperlinks among them
And the Problem is Real
Data on the Web (1 ★) is Not Enough

- Need a proper infrastructure for a real Web of Data
- Data is available on the Web
  - Accessible via standard Web technologies
- Data are interlinked over the Web
  - i.e., data can be integrated over the Web
- This is where Semantic Web technologies come in
Costs & Benefits of Web Data

• As a consumer/publisher
  • You can read it
  • You can print it
  • You can store it locally
  • You can export it to another system
  • You can modify it
  • You can share it (with conditions)
Costs & Benefits of ★★★Web Data

• Everything from one ★ plus:
  • You can directly process it with proprietary software to aggregate it, perform calculations, visualizations, etc.
  • You can export it into another structured format
Costs & Benefits of Web Data

- Everything with 2 ★★☆ plus:
  - You can manipulate the data in any way without software restrictions
Costs & Benefits of Web Data

- Everything with ★★★★★ plus:
  - You can link to it from anywhere
  - You can bookmark it
  - You can reuse parts of the data
  - You can combine the data with other data
  - You have complete control over the data items and can optimize their access
  - Other publishers can now link to your data making it
Costs & Benefits of Web Data

- Everything with ★★★★★ plus:
  - You can discover more (related) data while consuming the data
  - You can learn about the data schema
  - You can make your data discoverable
  - You increase the value of your data
Linked Data Principles

• Use URIs as names of things
  • Anything, not just documents
  • You are not your homepage
  • Information resources and non-information resources
• Use HTTP URIs
  • Globally unique names, distributed ownership
  • Allows people to look up those names
• Provide useful information in RDF
  • When someone looks up a URI
• Include RDF links to other URIs
  • To enable discovery of related information
RDF

- A data format for describing things and their interrelationships
- Standardized (XML)
- Easily parsed by machines
<table>
<thead>
<tr>
<th>Code</th>
<th>Title</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECOM6013: exam</td>
<td>E-commerce technologies</td>
<td>Prof. Bebo White</td>
</tr>
<tr>
<td>ECOM6029</td>
<td>E-business transformation</td>
<td>Prof. Ali Farhoodmand (HKU)</td>
</tr>
<tr>
<td>ECOM6030</td>
<td>Web 2.0 strategy and innovation</td>
<td>Prof. Amy Shuen (CEIBS)</td>
</tr>
<tr>
<td>ECOM6031: exam</td>
<td>Fundamentals of e-commerce security</td>
<td>Dr. KP Chow (HKU)</td>
</tr>
<tr>
<td>ECOM6035</td>
<td>Developing business models for digital media and online games</td>
<td>Mr. Peter Looms (University of Copenhagen)</td>
</tr>
<tr>
<td>ICOM6012: exam</td>
<td>Internet infrastructure technologies</td>
<td>Prof. Lawrence Yeung (HKU)</td>
</tr>
<tr>
<td>ICOM6043: exam</td>
<td>Information architecture</td>
<td>Prof. Renato Iannella (NETHA)</td>
</tr>
</tbody>
</table>

Session time: weekday: 6:45pm - 9:45pm; weekend: 9:30am - 12:30pm; 2:00pm - 5:00pm; 6:45pm - 9:45pm
Venues: Please refer to the module homepages

<table>
<thead>
<tr>
<th>MON</th>
<th>TUE</th>
<th>WED</th>
<th>THU</th>
<th>FRI</th>
<th>SAT</th>
<th>SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Sep</td>
<td>2-Sep</td>
<td>3-Sep</td>
<td>4-Sep</td>
<td>5-Sep</td>
<td>6-Sep</td>
<td>7-Sep</td>
</tr>
<tr>
<td>ICOM6012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8-Sep</td>
<td>9-Sep</td>
<td>10-Sep</td>
<td>11-Sep</td>
<td>12-Sep</td>
<td>13-Sep</td>
<td>14-Sep</td>
</tr>
<tr>
<td>Mid-Autumn Festival</td>
<td>ECOM6013</td>
<td></td>
<td></td>
<td></td>
<td>ECOM6013 (pm)</td>
<td>ECOM6013 (am+pm)</td>
</tr>
<tr>
<td>15-Sep</td>
<td>16-Sep</td>
<td>17-Sep</td>
<td>18-Sep</td>
<td>19-Sep</td>
<td>20-Sep</td>
<td>21-Sep</td>
</tr>
<tr>
<td>ECOM6013</td>
<td>ICOM6012</td>
<td>ECOM6013</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22-Sep</td>
<td>23-Sep</td>
<td>24-Sep</td>
<td>25-Sep</td>
<td>26-Sep</td>
<td>27-Sep</td>
<td>28-Sep</td>
</tr>
</tbody>
</table>
1. Uses **RDF Data Model**

   Subject: ECOM-ICOM  
   Predicate: organizes  
   Object: ECOM6013

   starts: 10.09.2014

   takesPlaceIn: Hong Kong

2. Is serialized in triples:

   ECOM-ICOM organizes ECOM6013
   ECOM6013 starts "20140910^^xsd:date .
   ECOM6013 takesPlaceIn Hong Kong.

3. Uses **Content-negotiation**

   Client
   - GET [vocabulary URI]
   - Accept: application/rdf+xml

   Server
   - 303 See Other
   - Location: [RDF content location]

   GET [RDF content location]
   - Accept: application/rdf+xml

   Server
   - 200 OK
   - <RDF>
Why Publish Linked Data? (Why Be 5-Star?)

- Ease of discovery
- Ease of consumption
  - Standards-based data sharing
- Reduced redundancy
- Added value
  - Build ecosystems around your data/content
Publishing Linked Open Data

• Identify and analyze your data

• Clean your data (?)

• Model your data (URI schema, vocabularies)

• Select and specify license(s)

• Convert your data to RDF

• Link your data to other data

• Publish and promote your Linked Open Data

• Watch others use it (become 5-star!)
Consuming Linked Open Data

- Specify concrete use cases
- Evaluate relevant data sources and data sets
- License clearing (check respective licenses)
- Create data consumption patterns
- Manage alignment, caching, and update mechanisms
- Create mash-ups, GUIs, services, and applications on top of the data
- Establish sustainable new partnerships
URI Scheme for US data.gov

URI Schemes

Submitted by jahendler on Fri, 2010-11-19 01:21

On the data.gov Web site, George Thomas and I have proposed that we need an effort to help develop a URI solution for the resolving of linked data on data.gov. We explain:

---

Data.gov now hosts a set of Resource Description Framework (RDF) documents containing triples created by converting a number of the Data.gov datasets into this format, making over 6.4 billion triples of open government data available to the community ...

The URI scheme chosen is a very simple one for the time being, designed to allow users to easily explore and extend the data. A proposal is being developed with RPI, one of the Data.gov community leaders, for a new encoding of datasets converted from CSV (and other formats) to RDF. We're looking forward to a design discussion to determine the best scheme for persistent and dereferenceable government URI naming with the international community and the World Wide Web Consortium to promote international standards for persistent government data (and metadata) on the World Wide Web.

---

The RPI proposal that is mentioned is discussed at: http://data-gov.tw.rpi.edu/wiki/A_Proposal_for_Governmental_Data_URIs

Active Group topics

- Status of NRC-Regulated Complex Materials Sites Undergoing Decommissioning
  | Last comment 2 days ago
- Annual 2008 Electric Power Industry Data
  | Last comment 2 days ago
- Alternative Transportation Fuels (ATF) and Alternative Fueled Vehicles (AFV) 2008
  | Last comment 2 days ago
- Annual 2008 Electric Generator Report (EIA-860)
  | Last comment 2 days ago
The Linking Open Data Project
The Linking Open Data Project

- Community project with W3C support
- Take existing open data sets
- Make them available on the Web in RDF
- Interlink them with other data sets
- Began in early 2007
The Linked Open Data Cloud

180 datasets, 20 billion RDF triples
Hi all,

we are happy to announce the release of DBpedia 2014.

The most important improvements of the new release compared to DBpedia 3.9 are:

1. the new release is based on updated Wikipedia dumps dating from April / May 2014 (the 3.9 release was based on dumps from March / April 2013), leading to an overall increase of the number of things described in the English edition from 4.26 to 4.58 million things.

2. the DBpedia ontology is enlarged and the number of infobox to ontology mappings has risen, leading to richer and cleaner data.

The English version of the DBpedia knowledge base currently describes **4.58 million things**, out of which 4.22 million are classified in a consistent ontology (http://wiki.dbpedia.org/Ontology2014), including 1,445,000 persons, 735,000 places (including 478,000 populated places), 411,000 creative works (including 123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (including 58,000 companies and 49,000 educational institutions), 251,000 species and 6,000 diseases.
The SLAC National Accelerator Laboratory, originally named Stanford Linear Accelerator Center, is a United States Department of Energy National Laboratory operated by Stanford University under the programmatic direction of the U.S. Department of Energy Office of Science.

- The SLAC National Accelerator Laboratory, originally named Stanford Linear Accelerator Center, is a United States Department of Energy National Laboratory operated by Stanford University under the programmatic direction of the U.S. Department of Energy Office of Science. The SLAC research program centers on experimental and theoretical research in elementary particle physics using electron beams and a broad program of research in atomic and solid-state physics, chemistry, biology, and medicine using synchrotron radiation.
- SLAC on vuonna 1962 perustettu Stanfordin yliopiston yhteydessä toimiva hiukkaskiihytyn laboratorio Kaliforniassa Yhdysvalloissa. SLAC keskityy alkeishiuukkasten kokeelliseen ja teoreettiseen tutkimukseen.
- Le Centre de l'accélérateur linéaire de Stanford (en anglais Stanford Linear Accelerator Center) est un laboratoire de physique dépendant du Département de l'Énergie des États-Unis et opéré par Université Stanford. Ses activités de recherche se concentrent sur la physique des particules théorique et expérimentale, et depuis quelques années s'ouvrent à la photonique au travers du projet LCLS. L'accélérateur de particules de 3,2 km de long situé sur le site est le plus long accélérateur linéaire au monde. Fichier:Stanford-linear-accelerator-usgs-ortho-kaminski-5900.jpg Vue aérienne de l'accélateurlinéaire]
We propose the following methodology for Linked Open Data publication:

1. **Publication & Applications**
   - Multiple Network Metrics
   - Network Metrics vs. University Metrics
   - ETL Process for General Data
   - ETL Process for Measurement Data

2. **Triplification & Linkage**
   - Domain Analysis
   - Ontology Engineering
   - Ontology Reuse
   - ontology Reuse Evaluation
   - Geonames Ontology
   - PingER LOD Ontology

3. **System Modeling and Results**
   - Understanding PingER Project's domain
   - Ontology Engineering
   - Ontology Reuse
   - Ontology Reuse Evaluation
   - Geonames Ontology
   - PingER LOD Ontology

4. **Research Design and Methodology**
   - ABSTRACT
   - CONCLUSIONS
   - FUTURE WORK
   - ACKNOWLEDGEMENT
   - REFERENCES

**ABSTRACT**

Most of the data published on the web is unstructured or does not follow a standard. It makes the data harder to be retrieved and interchanged between different data sources. Linked Open Data (LOD) technologies are applied in a scenario that deals with a large amount of computer network measurement data. As a result, we generated more structured data, hence easier to be retrieved, analyzed, and more interoperable. The challenges of processing large amount of data to transform it into a standard format (RDF); link it to other data sources; and analyze and visualize the transformed data are discussed. An ontology that aims to minimize the number of triples is proposed and a discussion on how ontologies may impact query performance is presented.

**CONCLUSIONS**

This work followed the methodology proposed to publish Linked Open Data applied in a real scenario that deals with big datasets about internet measurement. This methodology is based on:

- Domain analysis: understanding the domain and selecting which should be triples.
- Ontology engineering: reuse evaluation and number of triples minimization.
- Triplication project based on a parallel and distributed approach, linking to other data sources in the LOD cloud.

Publication: Enabling public access to both the data and the ontology in a standard, open, structured, and interoperable format, utilizing Semantic Web and LOD technologies.

**FUTURE WORK**

Utilizing complex SPARQL queries (those that are common in database with OLAP characteristics) on the PingER LOD database is still taking undesirable amount of time. Thus, in terms of query performance, more research is needed to find an efficient way of querying very large Triple Stores with OLAP characteristics.

**REFERENCES**


**ACKNOWLEDGEMENT**

This work was supported in part by the Department of Energy contract DE-AC02-76SF00515.

This work was supported in part by the Rio de Janeiro State Science Foundation (FAPERJ).
PingER Reuse Possibilities?

• What other parties might be interested in published PingER linked open data?
  • Network/Internet analysis researchers
  • Telecoms
  • IT managers at participating sites
  • “Digital Divide”/Education - “quality of service”
  • Emergency services
  • Etc., etc.

• What other parties might be interested in merging PingER linked open data?
The Power of Linked Open Data

• Enables Web-scale data publishing and discovery

• Everything is a potential resource

• Everything can be annotated

• Easy to extend and add new properties

• Easy to merge new RDF graphs and data ontologies

• Easy use and re-use on top of common schemas AND schema mapping

• Allows complex querying of multiple distributed data sources and systems
Why is Linked Open Data Important?

- Because in many cases it’s our data!
- Efficiency, reducing redundancy
- Promotes a digital society
- Opens the door to data innovation and discovery
- Holds the promise of creating from data
  - Knowledge
  - Wisdom
  - Benefit for all
Thanks for Your Attention!

Questions? Comments?
bebo@slac.stanford.edu
bebo@baiyongxue.com