The Emerging Web of Linked Data

Christian Bizer, Freie Universität Berlin
Outline

1. From a Web of Documents to a Web of Data
   - Web APIs and Linked Data

2. Linked Data Deployment on the Web
   - What data is out there?

3. Applications
   - What is being done with the data?

4. Next steps
   - What is still missing?
The Classic Web

Single Global Information Space

1. URLs as
   - globally unique IDs
   - retrieval mechanism

2. HTML as shared content format

3. Hyperlinks

Christian Bizer: The Web of Linked Data (26/07/2009)
Problem

As Web content is only loosely structured it is difficult for applications to do smart things with it.

Solution

Increase the structure of Web content.
Web APIs and Mashups
Web APIs and Mashups

Shortcomings

1. APIs provide proprietary interfaces
2. Mashups are based on a fixed set of data sources.
3. You can not set hyperlinks between data objects.

Christian Bizer: The Web of Linked Data (26/07/2009)
Web APIs slice the Web into Walled Gardens
Use Semantic Web technologies to
1. publish structured data on the Web,
2. set links between data from one data source
to data within other data sources.
1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful RDF information.
4. Include RDF statements that link to other URIs so that they can discover related things.

Tim Berners-Lee 2007

http://www.w3.org/DesignIssues/LinkedData.html
The RDF Data Model

```
foaf:Person rdf:type pd:cygri

foaf:name Richard Cyganiak

foaf:based_near dbpedia:Berlin
```
Data items are identified with HTTP URIs

\[ \text{pd:cygri} = \text{http://richard.cyganiak.de/foaf.rdf#cygri} \]
\[ \text{dbpedia:Berlin} = \text{http://dbpedia.org/resource/Berlin} \]
Resolving URIs over the Web

- pd:cygri rdf:type foaf:Person
  - foaf:name Richard Cyganiak
  - foaf:based_near dbpedia:Berlin
  - dp:population 3.405.259
  - skos:subject dp:Cities_in_Germany
Dereferencing URIs over the Web

- pd:cygri rdf:type foaf:Person
  - foaf:name Richard Cyganiak
  - foaf:based_near dbpedia:Berlin
    - skos:subject dbpedia:Muenchen
    - skos:subject dbpedia:Berlin
    - skos:subject dbpedia:Hamburg
    - skos:subject dp:Cities_in_Germany
  - dp:population 3.405.259

Christian Bizer: The Web of Linked Data (26/07/2009)
# Richard Cyganiak

**URI:** http://richard.cyganiak.de/foaf.rdf#cygri

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>event</td>
<td>...</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>G1 G2 G3 G4</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/cygri.rdf">http://richard.cyganiak.de/cygri.rdf</a></td>
<td>G2</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/foaf.rdf">http://richard.cyganiak.de/foaf.rdf</a></td>
<td>G3</td>
</tr>
<tr>
<td>nearest airport</td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td>phone</td>
<td>tel:+49-175-5630408</td>
<td>G1</td>
</tr>
<tr>
<td>sameAs</td>
<td>Richard Cyganiak</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td>...</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td>Berlin</td>
<td>G1</td>
</tr>
<tr>
<td>based_near</td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G1</td>
</tr>
<tr>
<td>currentProject</td>
<td><a href="http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs">http://page.mi.fu-berlin.de/~cyganiak/foaf.rdf#StatCvs</a></td>
<td>G3</td>
</tr>
<tr>
<td>currentProject</td>
<td><a href="http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq">http://www.wiwiss.fu-berlin.de/suhl/bizer#d2rq</a></td>
<td>G3</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="image.jpg" alt="Image" /></td>
<td>G4</td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
<td>G1</td>
</tr>
</tbody>
</table>
Berlin

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>population</td>
<td>3398888</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://dbpedia.org/City">http://dbpedia.org/City</a></td>
<td>G2</td>
</tr>
<tr>
<td>comment</td>
<td>Berlin is the capital city and one of the sixteen Federal States of Germany. It is the country's largest city in area and population, and the second most populous city in the European Union.</td>
<td>G2</td>
</tr>
<tr>
<td>comment</td>
<td>Berlin ist die deutsche Bundeshauptstadt und als Stadtstaat ein eigenständiges Land der Bundesrepublik Deutschland. Berlin ist die bevölkerungsreichste und flächengrößte Stadt Deutschlands und nach Einwohnern die zweitgrößte Stadt der EU.</td>
<td>G2</td>
</tr>
<tr>
<td>label</td>
<td>Berlin</td>
<td>G2</td>
</tr>
<tr>
<td>sameAs</td>
<td><a href="http://sws.geonames.org/2950159/">http://sws.geonames.org/2950159/</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Berlin">http://dbpedia.org/resource/category/Berlin</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Capitals_in_Europe">http://dbpedia.org/resource/category/Capitals_in_Europe</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Cities_in_Germany">http://dbpedia.org/resource/category/Cities_in_Germany</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/German_state_capitals">http://dbpedia.org/resource/category/German_state_capitals</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games">http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games</a></td>
<td>G2</td>
</tr>
<tr>
<td>subject</td>
<td><a href="http://dbpedia.org/resource/category/States_of_Germany">http://dbpedia.org/resource/category/States_of_Germany</a></td>
<td>G2</td>
</tr>
<tr>
<td>sourceURL</td>
<td>Berlin</td>
<td>G1</td>
</tr>
<tr>
<td>depiction</td>
<td><img src="http://en.wikipedia.org/wiki/Berlin" alt="Image of Berlin" /></td>
<td></td>
</tr>
<tr>
<td>birthplace</td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
</tbody>
</table>
Properties of the Web of Linked Data

- Anyone can publish data to the Web of Linked Data

- Entities are connected by links
  - creating a global data graph that spans data sources and enables the discovery of new data sources.

- Data is self-describing
  - If an application encounters data represented using an unfamiliar vocabulary, the application can resolve the URIs that identify vocabulary terms in order to find their RDFS or OWL definition.

- The Web of Data is open
  - meaning that applications can discover new data sources at run-time by following links.
2. Linked Data Deployment on the Web

- Is this real?
Grassroots community effort to

- publish existing open license datasets as Linked Data on the Web
- interlink things between different data sources
LOD Datasets on the Web: May 2007

- Over 500 million RDF triples
- Around 120,000 RDF links between data sources
LOD Datasets on the Web: September 2008

As of September 2008
LOD Datasets on the Web: March 2009

As of March 2009
 lod data set statistics as of July 2009

<table>
<thead>
<tr>
<th>Domain</th>
<th>No of Triples</th>
<th>% of Cloud</th>
<th>No of Links</th>
<th>% of Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media</td>
<td>698,000,000</td>
<td>10,4%</td>
<td>1,238,000</td>
<td>0,8%</td>
</tr>
<tr>
<td>Publications</td>
<td>212,000,000</td>
<td>3,2%</td>
<td>4,922,000</td>
<td>3,3%</td>
</tr>
<tr>
<td>Life Sciences</td>
<td>2,429,000,000</td>
<td>36,1%</td>
<td>133,199,000</td>
<td>89,4%</td>
</tr>
<tr>
<td>Geographic Data</td>
<td>3,097,000,000</td>
<td>46,0%</td>
<td>4,038,000</td>
<td>2,7%</td>
</tr>
<tr>
<td>User Generate Content</td>
<td>76,000,000</td>
<td>1,1%</td>
<td>1,559,000</td>
<td>1,0%</td>
</tr>
<tr>
<td>Cross-Domain</td>
<td>214,000,000</td>
<td>3,2%</td>
<td>3,992,000</td>
<td>2,7%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6,726,000,000</strong></td>
<td><strong>148,948,000</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

+ 2 billion triples from Data.gov published yesterday.
DBpedia is a community effort to
- extract structured information from Wikipedia
- make this information available on the Web under an open license
- interlink the DBpedia dataset with other open datasets on the Web

Contributors
- Freie Universität Berlin (Germany)
- Universität Leipzig (Germany)
- OpenLink Software (UK)
- Linking Open Data Community (W3C SWEO)
Extracting Structured Information from Wikipedia

http://en.wikipedia.org/wiki/Calgary

<http://dbpedia.org/resource/Calgary>

dbpedia:native_name "Calgary" ;
dbpedia:elevation "1048" ;
dbpedia:population_city "988193" ;
dbpedia:populationMetro "1079310" ;
mayor_name
$dbpedia:Dave_Bronconnier ;
governing_body
$dbpedia:Calgary_City_Council ;
...

- using a PHP extraction framework
- GPL license
The DBpedia Dataset

- Data about 2.6 million “things”
  - including at least
  - 213,000 persons
  - 328,000 places
  - 57,000 music albums
  - 36,000 films
  - 20,000 companies.

- Altogether 274 million pieces of information (RDF triples)
  - 609,000 links to images
  - 3,150,000 links to external web pages
  - 4,878,100 data links into external RDF datasets
Multi-Lingual Abstracts

The dataset contains a short and a long abstract for each concept.

**Short abstracts**

- **English**: 2,613,000
- **German**: 391,000
- **French**: 383,000
- **Dutch**: 284,000
- **Polish**: 256,000
- **Italian**: 286,000
- **Spanish**: 226,000
- **Japanese**: 199,000
- **Portuguese**: 246,000
- **Swedish**: 144,000
- **Chinese**: 101,000

**Property** | **Value**
---|---
abstract | 
English: 2,613,000
German: 391,000
French: 383,000
Dutch: 284,000
Polish: 256,000
Italian: 286,000
Spanish: 226,000
Japanese: 199,000
Portuguese: 246,000
Swedish: 144,000
Chinese: 101,000

Christian Bizer: The Web of Linked Data (26/07/2009)
1. Improving Wikipedia Search
2. Royalty-Free Data Source for other Applications
3. Nucleus for the Emerging Web of Data
Example Data Links

■ Out-Bound Link

<http://dbpedia.org/resource/Berlin> owl:sameAs
<http://sws.geonames.org/2950159> .

■ In-Bound Links

<http://richard.cyganiak.de/foaf.rdf#cygri> foaf:topic_interest

<http://blog.bizer.de/item1143> dc:subject
3. Linked Data Applications

What can I do with this?

Linked Data Browsers

Linked Data Mashups

Search Engines

DB

B

C

A

D

E

typed links

typed links

typed links

typed links
- Tabulator Browser (MIT, USA)
- Marbles (FU Berlin, DE)
- OpenLink RDF Browser (OpenLink, UK)
- Zitgist RDF Browser (Zitgist, USA)
- Humboldt (HP Labs, UK)
- Disco Hyperdata Browser (FU Berlin, DE)
- Fenfire (DERI, Ireland)
Domain-specific applications using Linked Data from the Web
DBpedia Mobile

- Displays Wikipedia data on a map
- Smushes the data with data from other sources
Web of Data Search Engines

- Falcons (IWS, China)
- Sig.ma (DERI, Ireland)
- Swoogle (UMBC, USA)
- VisiNav (DERI, Ireland)
- Watson (Open University, UK)
Falcons

Christian Bizer: The Web of Linked Data (26/07/2009)

Object Search  Concept Search

Beijing

Supports Boolean operators, quotes, and wildcard characters.

Beijing
Types: Capital, City
Labels: 北京” || Pekin || Пекин” || 北京市” || Pequim || Pechino || Beijing || Pékin” || Peking || Pekín
http://dbpedia.org/resource/Beijing - Described in 184 documents

Beijing
Types: Subject,
Labels: Beijing
http://ontoworld.org/wiki/Special:URIResolver/Beijing - Described in 11 documents

Beijing Guoan
Types: Club
Labels: Beijing Hyundai || 北京国安” || 北京国安足球俱乐部” || Beijing Guoan
http://dbpedia.org/resource/Beijing_Guoan - Described in 30 documents

Beijing
The declaration of this URI may be unauthorized.
Types: Capital City
Labels: Beijing
http://lonely.org/russia#Beijing - Described in 5 documents
<table>
<thead>
<tr>
<th>Breadcrumbs</th>
<th>Document Resource Document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ivan Herman</td>
<td><a href="http://www.ivan-herman.net/">http://www.ivan-herman.net/</a></td>
</tr>
<tr>
<td>Ivan’s private site</td>
<td><a href="http://ivan-herman.name/">http://ivan-herman.name/</a></td>
</tr>
<tr>
<td>open source</td>
<td><a href="http://www.advogato.org/person/connolly/">http://www.advogato.org/person/connolly/</a></td>
</tr>
<tr>
<td>Paul Downey</td>
<td><a href="http://blog.whatfettle.com/">http://blog.whatfettle.com/</a></td>
</tr>
</tbody>
</table>
Chris Bizer

picture:

given name: Chris [3,5,9,10,16]
family name: Bizer [3,5,9,10,16]
is creator of:

- DBpedia: A Nucleus for a Web of Open Data | Semantic Web Dog Food [6,18]
- The_TriQL_P_Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. [16]
- D2R Server - Publishing Relational Databases on the Semantic Web. [16]
- Named Graphs, Provenance and Trust [16]
- hide value • just this value • which sources • reject sources •
- RAP: RDF API for PHP [16]
- Fresnel: A Browser-Independent Presentation Vocabulary for RDF [16]
- NGA-L: Named Graphs API for Java [16]
What are the big players doing?

- Yahoo! and Google have started to crawl Linked Data in its RDFa serialization as well as Microformats.

- **Yahoo!**
  - provides access to crawled data through the Yahoo BOSS API
  - is using the data within Yahoo Search Monkey to make search results more useful and visually appealing.

- **Google**
  - uses crawled RDF data for its Social Graph API
  - is planning to / uses crawled data to enhance search results snippets for reviews and people.
Yahoo! Search Monkey

Dr. Seuss’ Horton Hears a Who


• Reviews: ★★★★★ (173)
• MPAA Rating: G
• Running Time: 1 hr. 28 min.
• Release Date: March 14th, 2008

acmemovies.com/hortonhearsawho - Cached
Connecting the classic Web and Linked Data

- Annotate Web documents with Linked Data URIs

\[
<\text{http://data.semanticweb.org/conference/eswc/2007/paper-69}>
\text{dc:subject <http://dbpedia.org/resource/Machine_learning> .}
\]

- (Semi-) Automated Annotation Services using Named Entity Recognition
  - Open Calais (Thomsons Reuters) for news
  - Zemanta (startup) for blog posts

- Goals
  - Connect everything.
  - Improve search by using Linked Data as background knowledge.
  - Display Web of Data content as info boxes next to news, blog posts.
Next steps

- More data is becoming available ....
  - US and UK government data
  - bibliographic data via Open Archives ORE
  - hot topic in current EU FP7 “Intelligent Information Management” call

- What is still missing?
Applications want an integrated view on all data that is available about an real-world entity!
Linked Data Fusion - Requirements

1. Schema Mapping: Map data into a single schema
   - so that data can be rendered and queried properly.

2. Identity resolution: Smush data from all sources about a single real-world entity
   - while keeping track of information provenance.

3. Conflict Resolution: Resolve inconsistencies in the data
   - by applying different trust heuristics.
There is a pay as you go data integration paradigm emerging on the Web of Data

- Publish data first using different schemata
- Maybe use common vocabularies
- Publish mappings to the Web afterwards

How to derive best-effort answers based on heterogeneous Web data and partial mappings?

- Alon Halevy, et al.: Web-scale Data Integration: You can only afford to Pay As You Go
There are no facts on the Web!

The Web is a social thing and everything on the Web is a claim.

Increasing amount of research in Semantic Web community on quality assessment, trust, data-cleansing.

Move the trust layer down in the Semantic Web Layer Cake

- Right above RDF and below OWL, SPARQL and RIF?
How do we build interfaces that operate over such large amounts of data?

- How to aggregate the data in a meaningful way?

What will be their interaction paradigm?

- Will the browser be something like a Web-Excel including drill-down?
- Will end-users notice that they are using Linked Data?

How to explain data provenance and data fusion?

- Tim Berner-Lee‘s „Oh, yeah?“ button.

What will Google and Yahoo do with the data?

- Will search engines turn into answer engines?
Hands on: How to publish Linked Data?

- Read the “How to Publish Linked Data on the Web” tutorial
  - http://www4.wiwi.fu-berlin.de/bizer/pub/LinkedDataTutorial/

- Publishing Tools
  - D2R Server: Publishes relational data bases as Linked Data and via SPARQL
  - Pubby: Linked Data wrapper that can be used together with any RDF store

- Link Generation Tools
  - Silk – Link Discovery Framework
  - ODDlinker

- Join the W3C Linking Open Data community
  - Wiki: http://esw.w3.org/topic/SweolG/TaskForces/CommunityProjects/LinkingOpenData
  - Mailing list: public-lod@w3.org
Thanks!

References

- Overview Article
  Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data – The Story So Far

- Linking Open Data Project Wiki
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

- Tutorial on How to Publish Linked Data on the Web
  http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/