The Emerging Web of Linked Data

Chris Bizer, Freie Universität Berlin
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
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

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

pd:cygri \(\text{rdf:type}\) foaf:Person

foaf:name \(\text{Richard Cyganiak}\)

foaf:based_near \(\text{dbpedia:Berlin}\)
Data items are identified with HTTP URIs

```
pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri
dbpedia:Berlin = 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 \( \text{rdf:type} \) foaf:Person

foaf:name \( \rightarrow \) Richard Cyganiak

foaf:based\_near \( \rightarrow \) dbpedia:Berlin

dp:population \( \rightarrow \) 3.405.259

skos:subject \( \rightarrow \) dbpedia:Berlin

skos:subject \( \rightarrow \) dbpedia:Hamburg

skos:subject \( \rightarrow \) dbpedia:Muenchen

skos:subject \( \rightarrow \) dp:Cities\_in\_Germany

Christian Bizer: The Web of Linked Data (26/07/2009)
<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>![Image of Richard Cyganiak]</td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
<td>G1</td>
</tr>
<tr>
<td>Property</td>
<td>Value</td>
<td>Sources</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<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="https://via.placeholder.com/150" alt="Image" /></td>
<td>G2</td>
</tr>
<tr>
<td>is 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

As of May 2007
Example RDF Links

- RDF links from DBpedia to other data sources

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

```html
<http://dbpedia.org/resource/Tim_Berners-Lee> owl:sameAs
```
As of September 2008
LOD Datasets on the Web: March 2009

As of March 2009
LOD Datasets on the Web: July 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></td>
<td><strong>148.948.000</strong></td>
<td></td>
</tr>
</tbody>
</table>

+ 2 billion triples from Data.gov published yesterday.
3. Applications

What can I do with this?

Linked Data Browsers

Linked Data Mashups

Search Engines

Thing

Thing

Thing

Thing

Thing

Thing

Thing

Thing

A

typed links

B

typed links

C

typed links

D

typed links

E

typed links
Linked Data Browsers

- 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, Irland)
Domain-specific applications using Linked Data from the Web
DBpedia Mobile

- Geospatial entry point into the Web of Data
- Starts with DBpedia, Revyu and Flickr data
Web of Data Search Engines

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

given name: Chris
family name: Bizer

is creator of: DBpedia: A Nucleus for a Web of Open Data | Semantic Web Dog Food
The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies.
D2R Server - Publishing Relational Databases on the Semantic Web.
Named Graphs, Provenance and Trust

hide value just this value which sources reject sources

RAP: RDF API for PHP
Fresnel: A Browser-Independent Presentation Vocabulary for RDF
NG4L - Named Graphs API for Java
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

Movie Details | Showtimes & Tickets | Trailers & Clips | Reviews

- 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
  

- (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

- What is still missing?
Applications want an integrated view on all data that is available about a real-world entity!
1. **Map data into a single schema**
   - so that data can be rendered and queried properly.

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

3. **Resolve inconsistencies in the data**
   - by applying different trust heuristics.
Data Quality and Trust

- There are no facts on the Web!
- The Web is a social thing and everything on the Web is a claim.
- Therefore we more research 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?
Reasoning with Linked Data

- Topic at recent Dagstuhl perspectives seminar
- Vocabulary term cherry-picking
- Retrieving ontology fragments from the Web
- Retrieving (partial) mappings from the Web
- Reasoning with large amounts of (inconsistent) Web data
Pay As You Go Data Integration

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
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.wiwiss.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/SweoIG/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/