Evolving the Web into a Global Data Space

Prof. Dr. Christian Bizer
Freie Universität Berlin
Germany
1. Linked Data
   • What is the vision and goal?

2. Topology of the Web of Data
   • What data is out there?

3. Global Data Integration
   • How to split the integration effort?

4. Challenges and Opportunities
Web APIs expose proprietary Interfaces
Mashups are based on a fixed set of data sources. Adding a new data source requires manual effort. No single global data space.
Alternative Approach: Linked Data

Extend the Web with a single global data space.

1. by using RDF to publish structured data on the Web
2. by setting links between data items within different data sources.
Linked Data Principles

Set of best practices for publishing structured data on the Web in accordance with the general architecture of the Web.

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.

Flexible, schema-less graph data model.
Entities are identified with HTTP URIs

HTTP URIs take the role of global primary keys.

pd:cygri = http://richard.cyganiak.de/foaf.rdf#cygri
dbpedia:Berlin = http://dbpedia.org/resource/Berlin
Resolving URIs over the Web

The HTTP protocol brings together identification and retrieval again.
Following Links deeper into the Web

Richard Cyganiak

- foaf:Person
- foaf:name
- foaf:based_near
- pd:cygri
- dp:population
- dbpedia:Berlin
- dbpedia:Hamburg
- dbpedia:Muenchen

3.405.259

skos:subject

dp:Cities_in_Germany
<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" alt="Richard Cyganiak" /></td>
<td>G1</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>3398888pis</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://dbpedia.org/Category">http://dbpedia.org/Category</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="image_url" alt="Image" /></td>
<td>G2</td>
</tr>
<tr>
<td>is birthplace of</td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
</tbody>
</table>
Properties of the Web of Linked Data

- **Global, distributed data space build on a simple set of standards**
  - RDF, URIs, HTTP

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

- **Provides for data-coexistence**
  - Everyone can publish data to the Web of Linked Data
  - Everyone can express their personal view on things
2. Topology of the Web of Data

- What data is out there?
Grassroots community effort to
- publish existing open license datasets as Linked Data on the Web
- interlink things between different data sources.
Over 500 million RDF triples
Around 120,000 RDF links between data sources
LOD Datasets on the Web: September 2008

As of September 2008
- Over 26.9 billion RDF triples
- Over 436 million RDF links between data sources
### Growth of the Web of Linked Data

<table>
<thead>
<tr>
<th>Year</th>
<th>Datasets</th>
<th>Triples</th>
<th>Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>12</td>
<td>500,000,000</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>45</td>
<td>2,000,000,000</td>
<td>300%</td>
</tr>
<tr>
<td>2009</td>
<td>95</td>
<td>6,726,000,000</td>
<td>236%</td>
</tr>
<tr>
<td>2010</td>
<td>203</td>
<td>26,930,509,703</td>
<td>300%</td>
</tr>
</tbody>
</table>

Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
The EU is starting to publish Linked Data (LOD2, LATC)

Various other national efforts

W3C eGovernment Interest Group
Uptake in the Libraries Community

**Institutions publishing Linked Data**
- Library of Congress (subject headings)
- German National Library (PND dataset and subject headings)
- Swedish National Library (Libris - catalog)
- Hungarian National Library (OPAC and Digital Library)
- Europeana Digital Library just released data about 4 million artifacts

**Growth of Library Linked Data (2009-2010): 1000%**

**W3C Library Linked Data Incubator Group**

**Goals:**
1. Integrate Library Catalogs on global scale.
2. Interconnect resources between repositories (by topic, by location, by historical period, by ...).
Uptake in Life Sciences

- W3C Linking Open Drug Data Effort
- Bio2RDF Project
- Allen Brain Atlas

**Goal:** Smoothly integrate internal and external data.
Uptake in the Media Industry

Goal: Drive traffic to websites via search engines

- Publish data as RDF/XML or RDFa

The Data

As of January 2010, The New York Times has published approximately 100,000 subject headings as linked open data under a CC BY licence. We include both RDF documents and a human-friendly HTML version. The table below gives a breakdown of the various tag types and mapping strategies on data.nytimes.com.

<table>
<thead>
<tr>
<th>Type</th>
<th>Manually Mapped Tags</th>
<th>Automatically Mapped Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>4,978</td>
<td>0</td>
</tr>
<tr>
<td>Organizations</td>
<td>1,419</td>
<td>1,692</td>
</tr>
<tr>
<td>Locations</td>
<td>1,910</td>
<td>0</td>
</tr>
</tbody>
</table>

---

Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
http://www.ckan.net/group/lodcloud

Basic Statistics (Nov 2010)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Data Sets</th>
<th>Triples</th>
<th>Percent</th>
<th>RDF Links</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-domain</td>
<td>20</td>
<td>1,999,085,950</td>
<td>7.42</td>
<td>29,105,638</td>
<td>7.36</td>
</tr>
<tr>
<td>Geographic</td>
<td>16</td>
<td>5,904,980,833</td>
<td>21.93</td>
<td>16,589,086</td>
<td>4.19</td>
</tr>
<tr>
<td>Government</td>
<td>25</td>
<td>11,613,525,437</td>
<td>43.12</td>
<td>17,658,869</td>
<td>4.46</td>
</tr>
<tr>
<td>Media</td>
<td>26</td>
<td>2,453,898,811</td>
<td>9.11</td>
<td>50,374,304</td>
<td>12.74</td>
</tr>
<tr>
<td>Libraries</td>
<td>67</td>
<td>2,237,435,732</td>
<td>8.31</td>
<td>77,951,898</td>
<td>19.71</td>
</tr>
<tr>
<td>Life sciences</td>
<td>42</td>
<td>2,664,119,184</td>
<td>9.89</td>
<td>200,417,873</td>
<td>50.67</td>
</tr>
<tr>
<td>User Content</td>
<td>7</td>
<td>57,463,756</td>
<td>0.21</td>
<td>3,402,228</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>203</strong></td>
<td><strong>26,930,509,703</strong></td>
<td><strong>0.21</strong></td>
<td><strong>395,499,896</strong></td>
<td><strong>0.86</strong></td>
</tr>
</tbody>
</table>

More statistics

http://www4.wiwiss.fu-berlin.de/lodcloud/state/
Linked Data Applications

What can I do with this?

Linked Data Browsers

Linked Data Mashups

Search Engines

Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
Linked Data Browsers

Provide for navigating between data sources and for exploring the data space.

- Tabulator Browser (MIT, USA)
- Marbles (FU Berlin, DE)
- OpenLink RDF Browser (OpenLink, UK)
- Zitgist RDF Browser (Zitgist, USA)
- Disco Hyperdata Browser (FU Berlin, DE)
- Fenfire (DERI, Irland)
Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
Web of Data Search Engines

Crawl the data space and provide best-effort query answers over crawled data.

- Falcons (IWS, China)
- Sig.ma (DERI, Ireland)
- Swoogle (UMBC, USA)
- VisiNav (DERI, Ireland)
- Watson (Open University, UK)
<table>
<thead>
<tr>
<th>given name:</th>
<th>Chris [3,5,9,10,16]</th>
</tr>
</thead>
<tbody>
<tr>
<td>family name:</td>
<td>Bizer [3,5,9,10,16]</td>
</tr>
<tr>
<td>is creator of:</td>
<td>&quot;DBpedia: A Nucleus for a Web of Open Data</td>
</tr>
<tr>
<td></td>
<td>The TriQL.P Browser: Filtering Information using Context-, Content- and Rating-Based Trust Policies. [16]</td>
</tr>
<tr>
<td></td>
<td>&quot;D2R Server - Publishing Relational Databases on the Semantic Web.&quot; [16]</td>
</tr>
<tr>
<td></td>
<td>Named Graphs, Provenance and Trust [16]</td>
</tr>
<tr>
<td></td>
<td>RAP: RDF API for PHP [16]</td>
</tr>
<tr>
<td></td>
<td>Fresnel: A Browser-Independent Presentation Vocabulary for RDF [16]</td>
</tr>
<tr>
<td></td>
<td>NG4J: Named Graphs API for Java [16]</td>
</tr>
</tbody>
</table>
What are the big players doing?
Open Graph Protocol

- allows site owners to determine how entities are described in Facebook
- relies on RDFa for encoding structured data in HTML pages
- available since April 2010
Schema.org

- allows site owners to provide data to enrich search results.
- relies on Microdata for encoding structured data in HTML pages
- available since June 2011
Search Engines turn into Answering Engines

Data snippets within search results

Answer to a fact query

Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
Publishing structured data becomes a SEO topic

- Usage of RDFa has increased 510% between March, 2009 and October, 2010
- 430 million webpages contain RDFa

Source: Yahoo

Topology of the Web of Data

Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
The Web of Data provides equal opportunities

Everybody can crawl the data.

- different from earlier approaches
  - like Google Base
  - and Google Fusion Tables
- just as on the classic Web
3. Global Data Integration

Applications hate heterogeneity!

The wild wild west

My little world
The Dataspace Vision

Alternative to classic data integration systems in order to cope with growing number of data sources.

- Properties of dataspaces
  - no upfront investment into a global schema
  - rely on pay-as-you-go data integration
  - give best effort answers to queries

Franklin, M., Halevy, A., and Maier, D.: From Databases to Dataspaces

Linked Data relies on the Pay-as-You-Go Idea

- for Identity Management
- for Schema/Vocabulary Management
You publish links pointing at other data sources

Somebody else publishes links pointing at your data source

Every link is an integration hint (semantic clue)
Effort Distribution between Publishers and Consumer

Consumer data mines identity links

Effort Distribution

Publishers or third parties provides identity links

-Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
Everyone can use whatever vocabularies she likes to publish data on the Web.

Or invest effort into reusing Common Vocabularies

- Friend-of-a-Friend for describing people and their social network
- SKOS for representing topic taxonomies
- GoodRelations provides terms for describing products and business entities
- SIOC for describing forums and blogs
- Organization Ontology for describing the structure of organizations
- Music Ontology for describing artists, albums, and performances
- Review Vocabulary provides terms for representing reviews

Many data source use mixture of common and proprietary vocabulary terms.
Publish Vocabulary Links on the Web

Vocabulary Link

<http://xmlns.com/foaf/0.1/Person>
owl:equiv
<http://dbpedia.org/ontology/Person> .

- Simple Mappings: RDFS, OWL
  - rdfs:subClassOf, rdfs:subPropertyOf
  - owl:equivalentClass, owl:equivalentProperty

- Complex Mappings: R2R
  - provides value transformation functions
  - structural transformations

Every vocabulary link is an integration hint.
Existing Vocabulary Links

Vocabulary links:
Vocabularies referencing "foaf" (119)

Vocabularies referenced by "mo" (17)

Source: Linked Open Vocabularies
http://labs.mondeca.com/dataset/lov
Effort Distribution between Publisher and Consumer

Consumer defines or data mines mappings

Effort Distribution

Publisher reuses vocabularies

Publisher or third party publishes mappings

Christian Bizer: Evolving the Web into a Global Data Space – BNCOD 2011, Manchester (12/7/2011)
The overall data integration effort is split between data publishers, third parties, and the data consumer.

- **Data Publisher**
  - publishes data as RDF
  - sets identity links
  - reuses terms or publishes mappings

- **Third Parties**
  - set identity links pointing at your data
  - publish mappings to the Web

- **Data Consumer**
  - has to do the rest
  - using manual effort as well as record linkage and schema matching techniques
4. Challenges and Opportunities
Things that are already there

■ Data Publication Tools
  ● D2R Server, Pubby, Triplify, Talis Plattform

■ Data Access Tools
  ● LDspider, SQUIN, Linked Data Client Libraries

■ RDF Stores
  ● RDF stores are getting better
  ● Example: Virtuoso 2 to 4 times slower than RDBMS on TCP-H in RDF
  ● See also: Berlin SPARQL Benchmark

■ Identity Resolution Tools
  ● SILK, LIMES

■ Data Translation Tools
  ● R2R Mapping Framework
Things that are still missing

1. More research on **data space profiling** is needed.
   - What is in the data space and how does the content change over time?

2. More research on **machine learning** mappings and identity resolution heuristics within the Web context.
   - Identity links make it easier to learn vocabulary links.
   - Vocabulary links make it easier to learn identity links.

3. More research on **data quality assessment and SPAM detection** is needed.

4. More research on **pay-as-you-go data integration** is needed.
   - How do human, community and machine contributions play together over time?
prototypes exist

lots of opportunities for vertical applications
  • Product Search, Job Search, Event Search, …
Opportunities

1. Analytical Applications
   - Web-scale OLAP
   - Global Data Mining

2. Linked Data as Data Integration Technology
   - within organizations. Example: BBC
   - for networks of companies. Example: W3C Linking Open Drug Data
**Hands-on: How to play around with the data?**

- **Download the Billion Triples Challenge Dataset**
  - 2 billion triples (20GB gzipped)
  - crawled from the public Web of Linked Data in May/June 2011

- **Download the Sindice Dump**
  - 12 billion triples (164GB gzipped, ~1,16TB uncompressed)
  - crawled from the public Web of Linked Data and
  - includes RDFa, Microformat, and wrapped API data

- **If you do something interesting with the data**
  - submit your results to the Semantic Web Challenge until October 1\(^{st}\)
References
