Topology of the Web of Data

Prof. Dr. Christian Bizer
Freie Universität Berlin
Germany
Slide from 2007: What does the Web offer us today?

- HTML
- DB

[Diagram showing the relationship between HTML and DB]

Christian Bizer: Topology of the Web of Data (04/30/2012)
Use the Web like a single, global database
More and more Websites publish Structured Data

Microformats

RDFa

Linked Data

Microdata
Research Prototypes: SigMa

Chris Bizer

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 TriQLP 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]
Industry Uptake 2011: Schema.org

- ask site owners to embed data to enrich search results
- 200+ Types: Event, organization, person, place, product, review
- Encoding: Microdata (or alternatively subset of RDFa)

Christian Bizer: Topology of the Web of Data (04/30/2012)
Usage of Schema.org Data

Data snippets within search results

The Fillmore - Western Addition/NOPA - San Francisco, CA
★★★★★ 752 reviews - Price range: $$
752 Reviews of The Fillmore "Last night we went to see Chris Isaak and it was our first time at the Fillmore. We could not have been any more delighted with...
www.yelp.com/biz/the-fillmore-san-francisco - United States - Cached - Similar

The Fillmore San Francisco - The Fillmore Schedule | Eventful
View The Fillmore’s upcoming event schedule and profile - San Francisco, CA. The Fillmore, also known as Fillmore Auditorium, is located in San ...
The Radiators - Farewell Tour! - 100th GAMH show!
3 NIGHTS! - An Evening With - Dark Star Orchestra
Bird by Bird - The Soft White Sixties - The Trophy Fire ...
eventful.com › San Francisco venues - Cached - Similar

Catherine Zeta-Jones date of birth — 25 September 1969 - Feedback
According to wikipedia.org, imdb.com, talktalk.co.uk and 4 others - Show sources
Google’s Knowledge Graph*

- describes more than 200 million entities, such as places, people, products ...
- consists of commercial third-party data and Web data
- will increasingly be used by Google to answer queries:

* Wall Street Journal: Google Gives Search a Refresh
03/14/2012
1. Embedded Data in HTML
   - Microformats
   - RDFa
   - Microdata
   - WebDataCommons.org

2. Linked Data
   - Sharing the data integration effort
   - The Web of Linked Data

3. Conclusions
   - Opportunities
   - Challenges
Microformats

- Small data islands within HTML pages
- Microformats effort dates back to 2003
- Small set of fixed formats
  - hcard: people, companies, organizations, and places
  - XFN: relationships between people
  - hCalendar: calendaring and events
  - hListing: small-ads; classifieds
  - hReview: reviews of products, businesses, events
- Shortcoming of Microformats
  - cannot represent any kind of data
- indexed by Google and Yahoo since 2009
RDFa

- serialization format for embedding RDF data into HTML pages
- proposed in 2004, W3C Recommendation in 2008
- can be used together with any vocabulary
- can assign URIs as global primary keys to entities

```html
<html xmlns="http://www.w3.org/1999/xhtml"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:foaf="http://xmlns.com/foaf/0.1/">

  ...<div about="http://example.com/Peter" typeof="foaf:Person">
    <span property="foaf:name">Peter Smith</span> knows
  </div>

  ...
</html>
```
Open Graph Protocol

- allows site owners to determine how entities are displayed inside Facebook
- relies on RDFa for encoding data in HTML pages
- available since April 2010
Microdata

- alternative technique for embedding structured data
- proposed in 2009 by WHATWG as part of HTML5 work
- tries to be simpler than RDFa (5 new attributes instead of 8)
- W3C currently tries to reconcile the two alternative proposals
- Schema.org initially chose Microdata as preferred serialization

```html
1 <div itemscope itemtype="http://schema.org/Person" itemid="http://example.com/Peter">
2   <span itemprop="name">Peter Smith</span>
3   <a href="http://example.com/Paula" itemprop="knows">Paula Jones</a>
4 </div>
```
Common Crawl is a non-profit foundation dedicated to building and maintaining an open crawl of the web, thereby enabling a new wave of innovation, education and research.
WebDataCommons.org

- extracts all Microformat, Microdata, RDFa data from the Common Craw and provides the extracted data for download

- Two extractions runs
  - 2009/2010 CC Corpus: 2.5 billion HTML pages (28.9 Terabyte compressed)
  - Feb 2012 CC Corpus: 1.4 billion HTML pages (20.9 Terabyte compressed)

- used 100 machines on Amazon EC2
  - approx. 3000 machine/hours (spot instances of type c1.xlarge) ➞ 550 EUR

- Jointed project of

Freie Universität Berlin

Karlsruhe Institute of Technology
HTML Pages containing structured Data

- 1.4 billion HTML pages parsed (Common Crawl, Feb 2012)
- 188 million pages contained Microformat, Microdata, RDFa

13% of the HTML pages contain structured data

- Size of extracted data set: 3.2 billion RDF quads

Christian Bizer: Topology of the Web of Data (04/30/2012)
Breakdown by Format (Feb 2012)

<table>
<thead>
<tr>
<th>Format</th>
<th>URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>html-rdfa</td>
<td>67,901,246</td>
</tr>
<tr>
<td>html-microdata</td>
<td>26,929,865</td>
</tr>
<tr>
<td>html-mf-geo</td>
<td>2,491,933</td>
</tr>
<tr>
<td>html-mf-hcalendar</td>
<td>1,506,379</td>
</tr>
<tr>
<td>html-mf-hcard</td>
<td>61,360,686</td>
</tr>
<tr>
<td>html-mf-hlisting</td>
<td>197,027</td>
</tr>
<tr>
<td>html-mf-hresume</td>
<td>20,762</td>
</tr>
<tr>
<td>html-mf-hreview</td>
<td>1,971,870</td>
</tr>
<tr>
<td>html-mf-species</td>
<td>14,033</td>
</tr>
<tr>
<td>html-mf-hrecipe</td>
<td>422,289</td>
</tr>
<tr>
<td>html-mf-xfn</td>
<td>26,004,925</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>188,821,015</strong></td>
</tr>
</tbody>
</table>
Size of the crawl: approximately 10 billion HTML pages
RDFa Topics (2012)

- Sample size: 49,370,729 instances RDFa from Common Crawl
- 150 classes and 400 properties with 1000+ instances
- Top Classes

1. gd:Breadcrumb (13,541,661 Entities)
2. foaf:Image (4,705,292 Entities)
3. gd:Organization (3,430,437 Entities)
4. foaf:Document (2,732,134 Entities)
5. skos:Concept (2,307,455 Entities)
6. gd:Review-aggregate (2,166,435 Entities)
7. sloc:UserAccount (1,150,720 Entities)
8. gd:Rating (1,055,997 Entities)
9. gd:Person (880,670 Entities)
10. sio:types:Comment (666,844 Entities)
11. gd:Product (619,493 Entities)
12. gd:Address (615,930 Entities)
13. gd:Review (540,537 Entities)
14. mo:Track (444,998 Entities)
15. gd:Geo (380,323 Entities)
16. mo:Release (238,262 Entities)
17. commerce:Business (197,305 Entities)
18. sio:types:BlogPost (177,031 Entities)
19. mo:SignalGroup (174,289 Entities)
20. mo:ReleaseEvent (139,118 Entities)

gd = Google’s Rich Snippet Vocabulary
**RDFa Properties (2012)**

- 400 properties with 1000+ instances
- Top Properties

<table>
<thead>
<tr>
<th>Rank</th>
<th>Property</th>
<th>Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gd:url</td>
<td>14,608,691</td>
</tr>
<tr>
<td>2</td>
<td>gd:title</td>
<td>13,368,007</td>
</tr>
<tr>
<td>3</td>
<td>rdfs:label</td>
<td>4,048,474</td>
</tr>
<tr>
<td>4</td>
<td>dcterms:title</td>
<td>3,745,593</td>
</tr>
<tr>
<td>5</td>
<td>gd:name</td>
<td>3,269,417</td>
</tr>
<tr>
<td>6</td>
<td>content:encoded</td>
<td>2,747,794</td>
</tr>
<tr>
<td>7</td>
<td>gd:locality</td>
<td>2,740,225</td>
</tr>
<tr>
<td>8</td>
<td>gd:reviewed</td>
<td>2,600,872</td>
</tr>
<tr>
<td>9</td>
<td>skos:prefLabel</td>
<td>2,551,861</td>
</tr>
<tr>
<td>10</td>
<td>gd:street-address</td>
<td>2,338,861</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>gd:votes</td>
<td>388,031</td>
</tr>
<tr>
<td>41</td>
<td>ogp:title</td>
<td>346,704</td>
</tr>
<tr>
<td>42</td>
<td>ogp:site_name</td>
<td>344,724</td>
</tr>
<tr>
<td>43</td>
<td>ogp:type</td>
<td>342,148</td>
</tr>
<tr>
<td>44</td>
<td>foaf:depiction</td>
<td>341,439</td>
</tr>
<tr>
<td>45</td>
<td>ogp:url</td>
<td>336,192</td>
</tr>
<tr>
<td>46</td>
<td>gd:category</td>
<td>330,674</td>
</tr>
<tr>
<td>47</td>
<td>ogp:image</td>
<td>325,807</td>
</tr>
<tr>
<td>48</td>
<td>gd:reviewer</td>
<td>315,621</td>
</tr>
<tr>
<td>49</td>
<td>mo:publication_of</td>
<td>293,983</td>
</tr>
<tr>
<td>50</td>
<td>mo:release_event</td>
<td>276,935</td>
</tr>
</tbody>
</table>

ogp = Facebook‘s Open Graph Protocol
Yahoo Crawl (2011)

- 12 billion pages, with 431 million pages containing RDFa

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Number of sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dublin Core (dc)</td>
<td>344,545</td>
</tr>
<tr>
<td>Open Graph Protocol (ogp)</td>
<td>177,761</td>
</tr>
<tr>
<td>Creative Commons (cc)</td>
<td>37,890</td>
</tr>
<tr>
<td>Google’s Rich Snippets Vocabulary (rich)</td>
<td>6,083</td>
</tr>
<tr>
<td>Friend-of-a-Friend (foaf)</td>
<td>2,545</td>
</tr>
<tr>
<td>Semantically-Interlinked Online Communities (sioc)</td>
<td>1,633</td>
</tr>
<tr>
<td>Electronic Business Cards (vcard)</td>
<td>1,349</td>
</tr>
<tr>
<td>Good Relations (goodrel)</td>
<td>488</td>
</tr>
<tr>
<td>Reviews (rev)</td>
<td>369</td>
</tr>
<tr>
<td>CommonTag (tag)</td>
<td>272</td>
</tr>
<tr>
<td>iCalendar Schema (ical)</td>
<td>62</td>
</tr>
</tbody>
</table>

Christian Bizer: Topology of the Web of Data (04/30/2012)
Microdata Topics (2012)

- Sample size: 90,526,013 Entities from the Common Crawl
- 182 classes and 690 properties with 1000+ instances

**Top Classes**

1. datavoc:Breadcrumb (18,528,472 Entities)
2. schema:VideoObject (10,760,983 Entities)
3. schema:Offer (6,608,047 Entities)
4. schema:PostalAddress (5,714,201 Entities)
5. schema:MusicRecording (2,054,647 Entities)
6. schema:AggregateRating (2,035,318 Entities)
7. schema:Product (1,811,496 Entities)
8. schema:Person (1,746,049 Entities)
9. datavoc:Offer (1,542,498 Entities)
10. schema:Article (1,243,972 Entities)
11. schema:WebPage (1,189,900 Entities)
12. datavoc:Rating (1,135,718 Entities)
13. schema:Review (1,016,285 Entities)
14. schema:Organization (1,011,754 Entities)
15. schema:Rating (872,688 Entities)
16. datavoc:Organization (861,558 Entities)
17. datavoc:Product (647,419 Entities)
18. datavoc:Person (564,921 Entities)
19. datavoc:Review-aggregate (539,642 Entities)
20. datavoc:Address (538,163 Entities)

datavoc = Google’s Rich Snippet Vocabulary
schema = Schema.org

Christian Bizer: Topology of the Web of Data (04/30/2012)
Instances per Class

- Microdata 02/2012
- RDFa 02/2012
- RDFa 2009/2010
- Microdata 2009/2010
RDFa and Microdata grow, but Microformats are still present

A rather small set of vocabularies is used

The content and the vocabularies are very focused towards the major consumers (Google, Yahoo, Bing, Facebook)

Providing structured data has come SEO topic

The data structures used are rather simplistic (mostly atomar entities)
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.
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
By following RDF links applications can

- navigate the global data graph
- discover new data sources
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


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

- for Identity Management
- for Schema/Vocabulary Management
Providing Integration Hints

- by publishing **Identity Links** on the Web

  
  **Identity Link**
  
  `<http://www4.wiwiss.fu-berlin.de/is-group/resource/persons/Person4>`
  `owl:sameAs`
  `<http://dblp.l3s.de/d2r/resource/authors/Christian_Bizer>`.

- you publish links pointing at other data sources.
- somebody else publishes links pointing at your data source.
Effort Distribution between Publisher and Consumer

Consumer data mines identity links

Effort Distribution

Publishers or third parties provides identity links

Application Layer
- Application Code

Data Access, Integration and Storage Layer
- Web Data Access Module
- Vocabulary Mapping Module
- Identity Resolution Module
- Quality Evaluation Module
- Integrated Web Data

Web of Linked Data

Publication Layer
- LD Wrapper
- Database A
- LD Wrapper
- Database B
- RDFa
- Legacy App C
- RDF/XML

SPARQL

HTTP

Christian Bizer: Topology of the Web of Data (04/30/2012)
Providing Integration Hints

- by publishing **Vocabulary Links** on the Web

Vocabulary Link

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

- **Terms for expressing Correspondences**
  - owl:equivalentClass, owl:equivalentProperty
  - rdfs:subClassOf, rdfs:subPropertyOf
Effort Distribution between Publisher and Consumer

Consumer defines or data mines mappings

Publisher reuses vocabularies

Publisher or third party publishes mappings

Application Layer

Application Code

Data Access, Integration and Storage Layer

Web Data Access Module → Vocabulary Mapping Module → Identity Resolution Module → Quality Evaluation Module → Integrated Web Data

Web of Linked Data

Publication Layer

LD Wrapper → Database A
LD Wrapper → Database B
RDFa → Legacy App C
RDF/XML → HTTP
The overall data integration effort is split between the data publisher, the data consumer and third parties.

- **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 record linkage and schema matching techniques
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
31.6 billion RDF triples
503 million RDF links
## Distribution by Topical Domain (Nov 2011)

<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>Media</td>
<td>25</td>
<td>1,841,852,061</td>
<td>5.82%</td>
<td>50,440,705</td>
<td>10.01%</td>
</tr>
<tr>
<td>Geographic</td>
<td>31</td>
<td>6,145,532,484</td>
<td>19.43%</td>
<td>35,812,328</td>
<td>7.11%</td>
</tr>
<tr>
<td>Government</td>
<td>49</td>
<td>13,315,009,400</td>
<td>42.09%</td>
<td>19,343,519</td>
<td>3.84%</td>
</tr>
<tr>
<td>Publications</td>
<td>87</td>
<td>2,950,720,693</td>
<td>9.33%</td>
<td>139,925,218</td>
<td>27.76%</td>
</tr>
<tr>
<td>Cross-domain</td>
<td>41</td>
<td>4,184,635,715</td>
<td>13.23%</td>
<td>63,183,065</td>
<td>12.54%</td>
</tr>
<tr>
<td>Life sciences</td>
<td>41</td>
<td>3,036,336,004</td>
<td>9.60%</td>
<td>191,844,090</td>
<td>38.06%</td>
</tr>
<tr>
<td>User content</td>
<td>20</td>
<td>134,127,413</td>
<td>0.42%</td>
<td>3,449,143</td>
<td>0.68%</td>
</tr>
<tr>
<td><strong>SUM</strong></td>
<td><strong>295</strong></td>
<td><strong>31,634,213,770</strong></td>
<td><strong>503,998,829</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: State of the LOD Cloud  
http://www4.wiwiss.fu-berlin.de/lodcloud/state/
Vocabulary Usage (Nov 2011)

- Only proprietary vocabularies: 104 (35.25 %) of the 295 sources
- Terms from non-proprietary vocabularies: 191 (64.75 %) of the 295 sources

Common Vocabularies

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc</td>
<td>92</td>
<td>31.19 %</td>
</tr>
<tr>
<td>foaf</td>
<td>81</td>
<td>27.46 %</td>
</tr>
<tr>
<td>skos</td>
<td>58</td>
<td>19.66 %</td>
</tr>
<tr>
<td>geo</td>
<td>25</td>
<td>8.47 %</td>
</tr>
<tr>
<td>akt</td>
<td>17</td>
<td>5.76 %</td>
</tr>
<tr>
<td>bibo</td>
<td>14</td>
<td>4.75 %</td>
</tr>
<tr>
<td>mo</td>
<td>13</td>
<td>4.41 %</td>
</tr>
<tr>
<td>vcard</td>
<td>10</td>
<td>3.39 %</td>
</tr>
<tr>
<td>sioc</td>
<td>10</td>
<td>3.39 %</td>
</tr>
<tr>
<td>cc</td>
<td>8</td>
<td>2.71 %</td>
</tr>
</tbody>
</table>

Source: State of the LOD Cloud http://www4.wiwiss.fu-berlin.de/lodcloud/state/
Deployment of Vocabulary Links

Vocabulary links:
Vocabularies referencing "foaf" (119)

Vocabularies referenced by "mo" (17)

Uptake in the Government Domain

- 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

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

- **W3C Library Linked Data Incubator Group**
Conclusion: Web of Linked Data

Compared to Microformats, Microdata, RDFa

- number of data providers is significantly lower
- wider range of topics covered
- wider range of common and proprietary vocabularies used
- more complex data structures
- emphasis on setting RDF Links between sources
Conclusion: Topology of the Web of Data
3. Opportunities and Challenges
The Web of Data provides equal Opportunities

Everybody can crawl the data.

- different from alternative approaches
  - like Google Base
  - like Facebook
  - like Google Fusion Tables

- just as on the classic Web

The haystack is there, so let's look for the needle!
Search Engines turn into Answering Engines

Ivan Herman
http://www.ivan-herman.net/
Document Resource Document

Ivan's private pito
http://ivan-herman.name/
RSS1.0 News Channel Document Resource

open source
http://www.advogato.org/person/connolly/
RSS1.0 News Channel Document organization
Advogato blog for connolly

Paul Downey
http://blog.whattelse.com/
Document Resource Document Whattelse, marras?

Movies for San Francisco, CA

The Hunger Games
2hr 22min PG-13 Action Trailer

21 Jump Street
3hr 49min R Action Trailer

Dr. Seuss' The Lorax
43 reviews 1hr 35min PG Animation Trailer

Dr. Seuss' The ...
43 reviews 1hr 35min PG Animation Trailer

John Carter
11 reviews 2hr 19min PG-13 Action Trailer

Act of Valor
42 reviews 1hr 51min R Action

+ Show more movies
Applications hate heterogeneity and low quality data!

The wild wild west

My little world
Things that require more work

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 data quality assessment and SPAM detection is needed.

3. More research on 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.

4. More research on pay-as-you-go data integration is needed.
   - How do human, community and machine contributions play together over time?
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 Web Data Commons Dump**
  - 3 billion triples (49 GB, gzipped)
  - RDFa, Microdata, Microformat data crawled February 2012

- **Download the Sindice Dump**
  - 12 billion triples (164GB gzipped, ~1,16TB uncompressed)
  - Linked Data, RDFa, Microdata, Microformat crawled 2009-2011
Thanks!

References

- Statistics on HTML-embedded data: http://webdatacommons.org
- Statistics on Linked Data: http://www4.wiwiss.fu-berlin.de/lodcloud/state/