Turning the Web into a Database

What does the Web offer us today?

Christian Bizer: Turning the Web into a Database (14.2.2007)
What do we actually want?

Outline

1. Development of the Web
2. Overview of some Base Technologies
3. Integrating Relational Databases into the Web
4. Summary and Outlook
Christian Bizer: Turning the Web into a Database (14.2.2007)

Development of the Web

- World Wide Web
- Web 2.0
- Web of Data
- Web 3.0
- Semantic Web

1995  2005  ?

Current Stage: Web 2.0

1. Read-and-Write Web
   - User-generated Content, Weblogs, Social Networks

2. Introduction of more structured content formats
   - Microformats:
     - Small data islands within HTML pages (hCard, hCalendar, hReview)
   - Really Simple Syndication (RSS):
     - Adds explicit timestamp to Web content

Increased structure enables more sophisticated information processing!
Really Simple Syndication (RSS)

```xml
<?xml version="1.0"?>
<rss version="2.0">
  <channel>
    <title>CNN.com</title>
    <language>en-us</language>
    <item>
      <title>Bush: Iraq plan is 'best chance'</title>
      <description>President George W. Bush says ...</description>
      <pubDate>Wed, 24 Jan 2007 07:34:44 EST</pubDate>
    </item>
    <item>
      <title>Rape charge for Israeli president</title>
      <description>Israel's attorney general has ...</description>
      <pubDate>Wed, 24 Jan 2007 12:44:21 EST</pubDate>
    </item>
    ...
  </channel>
</rss>
```

Use Case: Trend Scouting

- Crawler 63.2 million RSS feeds
- Provides keyword search and trend analysis
Use Case: Financial Information Analysis

Deutsche Telekom (ISIN: DE0005557508)

- Diploma Thesis
- Aggregates RSS feeds from:
  - 296 analyst houses
  - 700 newspapers
  - 1 feed about insider trades

Structure enables Information Processing

- Tim Berners-Lee (MIT/W3C)

Basic Ideas
- Publish pure data in addition to HTML pages on the Web
- Set links between data items within different data sources

Realization
- Pushed by W3C with several standards
- World wide research effort
- Industry is getting involved (Oracle, HP, IBM)

Base Technologies
- Resource Description Framework (RDF)
- Query Language for RDF (SPARQL)
Resource Description Framework (RDF)

- RDF provides a simple data model for publishing data on the Web.
- Information is represented in the form of triples.

**Triple: Chris knows Richard**

- **Subject:** http://www.bizer.de#chris
- **Predicate:** http://xmlns.com/foaf/0.1/knows
- **Object:** http://richard.cyganiak.de/foaf.rdf#cygri

**Triple: Chris was born 1973-09-20**

- **Subject:** http://www.bizer.de#chris
- **Predicate:** http://www.w3.org/2001/vcard-rdf/3.0#BDAY
- **Object:** "1973-09-20"^^http://www.w3.org/2001/XMLSchema#date

Every Triple is a Hyperlink

GET /city/Berlin HTTP/1.0
Accept: application/rdf+xml
Every Triple is a Hyperlink

rc:cygri rdf:type foaf:Person
foaf:name Richard Cyganiak
foaf:based_near dp:city/Berlin
dp:population 3.398.888
dp:city/Berlin
skos:subject dp: Cities_in_Germany

dp:city/Berlin
skos:subject dp: Cities_in_Germany

dp:city/Hamburg skos:subject dp: Cities_in_Germany
dp:city/Muenchen skos:subject dp: Cities_in_Germany
dp:district geo:Schwabing
dp:district geo:Freimann

dp:city/Berlin
skos:subject dp: Cities_in_Germany

dp:city/Hamburg skos:subject dp: Cities_in_Germany
dp:city/Muenchen skos:subject dp: Cities_in_Germany
dp:district geo:Schwabing
dp:district geo:Freimann

Christian Bizer: Turning the Web into a Database (14.2.2007)
### Richard Cyganik

**URL:** [http://richard.cyganiak.de/toaf.rdf#rgni](http://richard.cyganiak.de/toaf.rdf#rgni)

<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/toaf.rdf#rgni">http://richard.cyganiak.de/toaf.rdf#rgni</a></td>
<td>G2</td>
</tr>
<tr>
<td>seeAlso</td>
<td><a href="http://richard.cyganiak.de/toaf.rdf#rgni">http://richard.cyganiak.de/toaf.rdf#rgni</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 Cyganik #</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/toaf.rdf#StatCys">http://page.mi.fu-berlin.de/cyganiak/toaf.rdf#StatCys</a> #</td>
<td>G3</td>
</tr>
<tr>
<td>depiction</td>
<td><a href="http://www.w3cws.ts-berlin.de/shell/10x10/b2d.png">http://www.w3cws.ts-berlin.de/shell/10x10/b2d.png</a> #</td>
<td>G3 G4</td>
</tr>
<tr>
<td>gender</td>
<td>male</td>
<td>G1</td>
</tr>
</tbody>
</table>

### Berlin

**URL:** [http://dbpedia.org/resource/city/Berlin](http://dbpedia.org/resource/city/Berlin)

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>population</td>
<td>3396888</td>
<td>G2</td>
</tr>
<tr>
<td>type</td>
<td><a href="http://dbpedia.org/Category:City">http://dbpedia.org/Category: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 bevölkerungsrächtigste und flächengrößte Stadt Deutschlands und nach Einwohner 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/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/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/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><a href="http://www.bild.de/Bild">http://www.bild.de/Bild</a> Архив/2012/06/berlin_galerie_02_12.jpg #</td>
<td>G2</td>
</tr>
<tr>
<td>isBirthplaceOf</td>
<td>Adolf von Baeyer #</td>
<td>G2</td>
</tr>
</tbody>
</table>
### The SPARQL Query Language

- W3C standard query language for RDF data
- Example query

```
SELECT ?name ?birth
WHERE { ?person foaf:name ?name .
  FILTER (?birth > "1970-01-01"^^xsd:date) }
```

<table>
<thead>
<tr>
<th>?name</th>
<th>?birth</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Chris&quot;</td>
<td>&quot;1973-09-20&quot;^^xsd:date</td>
</tr>
<tr>
<td>&quot;Andy&quot;</td>
<td>&quot;1980-03-12&quot;^^xsd:date</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

### What to query?

1. Querying a Single Data Source
2. Querying Multiple Data Sources
3. Querying the complete Semantic Web
Querying a Single Data Source

- Data source: dbpedia.org
- 19 million triples extracted from Wikipedia info boxes.

Example Query

Give me all persons from Wikipedia that were born in Berlin before 1900.

```sql
SELECT ?name ?birth ?description
WHERE {
  ?person foaf:name ?name .
  FILTER (?birth < "1900-01-01"^^xsd:date )
}
```

SPARQL Explorer for http://dbpedia.org/sparql

Sparql results:

- Adolfo_de_Austria[@en] "1811-10-29"xsd:dateTime
  *Heinrich Wilhelm Adolphi Prince of Prussia (October 29, 1811 in Berlin – June 8, 1873 in Karlsbad) was a Prussian naval theorist and admiral. As a son of Prince William, the younger brother of King Friedrich William III, Adalbert was instrumental during the Revolutions of 1848 in founding the first united German fleet. During the 1890s he helped establish the Prussian Navy.*

- Adolf_von_Bayer[@en] "1835-10-31"xsd:dateTime
  *Johann Friedrich Wilhelm Adolf von Bayer (October 31, 1826 - August 20, 1917) was a German chemist who synthesized indigo, and was the 1906 recipient of the Nobel Prize in Chemistry. Born in Berlin, he initially studied mathematics and physics at Berlin University before moving to Heidelberg to study chemistry with Robert Bunsen. There he worked primarily in August Kekulé's laboratory, earning his doctorate from Berlin in 1859.*

- Albert_Lortzing[@en] "1801-10-23"xsd:dateTime
  *Gustaf Albert Lortzing (October 23, 1801 - January 21, 1855) was a German composer.*

- Alexander_Gottlieb_Baumgarten[@en] "1714-07-17"xsd:dateTime
  *Alexander Gottlieb Baumgarten (July 17, 1714 Andechs; May 26, 1762) was a German philosopher. He was a follower of Leibniz and Christian Wolff, and gave the term aesthetics its modern meaning.*

- Alexander_Mitscherlich[@en] "1836-05-26"xsd:dateTime
  *This article is about the chemist. Go to Alexander Mitscherlich (Psychologist) for the psychologist.*
Querying Multiple Data Sources

Three Options:
1. Virtual Integration
   - Query Routing: Split query and ask data sources for the things that they can answer.
   - HU Berlin: DARQ
   - Complicated and slow.
2. Materialized Integration
   - Use Links between data items to crawl all data into a single repository.
   - Crawlers under development: Zitgist, OpenLink USA and SWSE, DERI Ireland
   - Fast, but requires huge RDF repositories.
3. Materialization On-the-Fly
   - Crawl only data that is needed while answering the query.
   - FU Berlin: Semantic Web Client Library
   - Works, but is really slow.

Querying the complete Semantic Web

Two Options:
1. Materialized Integration
   - Use Links between data items to crawl all data into a single repository.
   - Worked for HTML (Google) and RSS (Technorati)
2. Materialization On-the-Fly
   - Crawl only data that is needed while answering the query.
   - Works, but is really slow.
Now we have a data model and a query language.
What is still needed is an RDF view on the complete Web.

Integrating Relational Databases into the Semantic Web

What is needed?
- A way to map data between the relational data model and RDF
- Some software to make the resulting RDF data Web-accessible

We developed two artifacts:
- The D2RQ Mapping Language
- D2R Server
The mapping has to bridge:

- **Data Model Heterogeneity**
  - relations versus a graph
  - primary keys versus URIs

- **Schema Heterogeneity**
  - name versus given name, surname

- **Syntactic Heterogeneity**
  - timestamp versus “2007-02-14”^^xsd:data
**Value Correspondences**

- **D2RQ** is a language for writing these correspondences down.

---

**Example D2RQ Mapping**

- **ClassMap People**
  - PropertyBridge refersTo
  - PropertyBridge foaf:name
  - PropertyBridge foaf:mbox
  - PropertyBridge dc:author
  - PropertyBridge rdf:type

- **ClassMap Papers**
  - PropertyBridge dc:title
  - PropertyBridge swc:conference
  - PropertyBridge rdf:type
### Class Map

**People**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Chris</td>
<td><a href="mailto:chris@bizer.de">chris@bizer.de</a></td>
</tr>
</tbody>
</table>

:PeopleClassMap a d2rq:ClassMap;
d2rq:class foaf:Person;
d2rq:uriPattern
"http://example.org/person@@People.ID@@".

### Property Bridge

**People**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Chris</td>
<td><a href="mailto:chris@bizer.de">chris@bizer.de</a></td>
</tr>
</tbody>
</table>

:PeopleEmailProperty a d2rq:PropertyBridge;
d2rq:belongsToClassMap :PeopleClassMap;
d2rq:property foaf:mbox;
d2rq:uriPattern "mailto:@@People.email@@".
### Joins

<table>
<thead>
<tr>
<th>People</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>name</td>
</tr>
<tr>
<td>12</td>
<td>Chris</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rel_People_Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>PersonID</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

:PeoplePaperRelation a d2rq:PropertyBridge;
  d2rq:belongsToClassMap :PeopleClassMap;
  d2rq:property dc:author;
  d2rq:refersToClassMap :PapersClassMap;
  d2rq:join "People.ID=Rel_People_Papers.PersonID";
  d2rq:join "Rel_People_Papers.PersonID=Papers.ID";

### D2R Server

- open source project (GNU General Public License)
- around 1500 downloads (138 in January 2007)
### Example Rewriting from SPARQL to SQL

```
SELECT ?person ?mbox
WHERE { ?person foaf:name "Chris" .
       ?person foaf:mbox ?mbox . }
```

```
SELECT People.ID, People.email
FROM People
WHERE People.name = "Chris";
```

### Usage of D2RQ and D2R Server

- **Public D2R Servers on the Web**
  - dbpedia, FU Berlin and Universität Leipzig
  - DBLP Bibliography, FU Berlin
  - Information Systems Group, FU Berlin
  - Roller, Weblog Server, Sun Microsystems
  - Images of Fruitfly Embryogenesis, Berkeley Drosophila Genome Project

- **OEM Distribution**
  - Part of the TopBraid Composer Ontology Editor
  - TopBraid is a startup company in Mountain View, California
  - Customers: NASA, US Military

Christian Bizer: Turning the Web into a Database (14.2.2007)
Publications and Outreach

- **Publications**

- **Outreach**
  - Bob DuCharme: Relational Database Integration with RDF/OWL. XML 2006 Conference, Boston, USA, December 2006
  - Alois Reibauer: Tool-Workshop Triple Stores, SPARQL und D2RQ. Semantic Web School Austria, Wien, January 2007 (340 € excl. USt.)

Summary

- Gave an overview about the development of the Web towards a database
- Introduced RDF and SPARQL
- Described the D2RQ Mapping Language
- Described D2R Server
Outlook

- Web 2.0 kick-started Semantic Web roll-out.
- There are more and more RDF data sources on the Web.
  - FOAF, DOAP, SIOC Communities
  - dbpedia
  - geonames
  - RDF Book Mashup
  - WordNet
  - MusicBrainz
- Semantic Web Search Engines are getting into place.
  - Zitgist, Openlink USA
  - SWSE, DERI Ireland
  - Swoogle, University of Maryland USA
- Things will stay exiting in the years to come.

Thanks!

This slides are online at
http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/disp-bizer.pdf