Linked Data, DBpedia and D2R Server

Building blocks for the Emerging Web of Data

Dr. Christian Bizer
Freie Universität Berlin

Outline

1. Approaches to Realize the Web of Data
2. Some Building Blocks
   1. DBpedia
   2. W3C Linking Open Data Project
   3. D2RQ and D2R Server
3. Discussion
What does the Web offer us today?

What do we actually want?

Use the Web like a single, global database
## Approaches to Realize the Web of Data

1. **Google Base**
2. **Freebase**
3. **Semantic Web**

### 1. Google Base

- Aims to be THE database for all data in the world
- **Flexible item-based data model**
  - Items have types and properties which can be defined by users
- **Data is uploaded by many data providers**
  - Lots of sales offers, event data, but also recipes
- **Royalty-free access via**
  - Graphical user interface
  - Application programming interface (API)
- **Google interlinks the data with other datasets they crawl or own**
Google Base User Interface

**2. Freebase**

<table>
<thead>
<tr>
<th>Pulp Fiction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td>Pulp Fiction is a highly acclaimed, Oscar-winning 1994 film directed by Quentin Tarantino, who co-wrote the screenplay with Roger Avary. The film is structured around a fragmented...</td>
</tr>
</tbody>
</table>
3. Semantic Web

- Tim Berners-Lee (MIT/W3C) outlined the Semantic Web vision, twice.
  - 2001, Scientific American article
    - Formal ontologies, hyper-intelligent agents, lots of AI
  - 2007, Linked Data web architecture note
    - Go back to the basic Web architecture
    - Aim at a data web that is useful in the short term
- Basic Ideas of Linked Data
  - Publish pure data in addition to HTML pages on the Web
  - Set links between data items within different data sources

Linked Data Principles

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
Every Triple is a Hyperlink!

GET /city/Berlin HTTP/1.0
Accept: application/rdf+xml

Dereferencing URIs over the Web
Dereferencing URIs over the Web

- **rc:cygri**\(\rightarrow\) **foaf:Person**
  - **foaf:name** → Richard Cyganiak
  - **foaf:based_near** → **dp:city/Berlin**
    - **dp:population** → 3,405,259
    - **skos:subject** → **dp:Cities_in_Germany**

- **dp:city/Berlin**
  - **foaf:based_near** → **dp:city/Hamburg**
    - **skos:subject** → **dp:Cities_in_Germany**
  - **skos:subject** → **dp:city/Muenchen**

Christian Bizer: Linked Data, DBpedia and D2R Server (26.06.2007)
Browsing the Semantic Web

- Tabulator Browser (MIT, USA)
- Disco Hyperdata Browser (FU Berlin)
- OpenLink RDF Browser (OpenLink, UK)
Querying the Semantic Web

Three Options:

1. **Virtual Integration**
   - **Query Routing**: Split SPARQL query and ask different data sources for the things that they might be able to answer.
   - **HU Berlin: DARQ**
   - Needs proper data source descriptions.
   - Complicated and slow.

2. **Materialized Integration**
   - Use Links between data items to crawl all data into a single repository.
   - Fast, but requires huge RDF repositories.
   - Worked for HTML, worked for RSS, so why not for RDF?

3. **Materialization On-the-Fly**
   - Crawl only data that is needed while answering the query.
   - **FU Berlin: Semantic Web Client Library**
   - **University of London: SWIC**
   - Works, but is really slow.
### Semantic Web Search Engines

- Currently under development
  - Zitgist (Zitgist, USA)
  - SWSE (DERI, Ireland)
  - Watson (Open University, UK)
  - Swoogle (UMBC, USA)
- Crawl RDF data by following RDF Links
- Can offer sophisticated query capabilities
- Can offer nice user interfaces
- Hopefully get better this year!

### 2. Some Building Blocks

1. DBpedia – Extracting Structured Information from Wikipedia
2. W3C Linking Open Data Community Project
3. D2R Server – Publishing Relational Databases on the Web
2.1 DBpedia

DBpedia.org is a community effort to
- extract structured information from Wikipedia
- make this information available on the Web as under an open license

Contributors
- Freie Universität Berlin (Germany)
- Universität Leipzig (Germany)
- OpenLink Software (UK)

Extracting Structured Information from Wikipedia

Wikipedia consists of
- 6.9 million articles
- in 251 languages
- monthly growth-rate: 4%

Wikipedia articles contain structured information
- infoboxes which use a template mechanism
- images depicting the article’s topic
- categorization of the article
- links to external webpages
- intra-wiki links to other articles
- inter-language links to articles about the same topic in different languages
Extracting Infobox Data

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

<http://dbpedia.org/resource/Calgary>
dbpedia:native_name "Calgary" ;
dbpedia:altitude "1048" ;
dbpedia:population_city "988193" ;
dbpedia:population_metro "1079310" ;
mayor_name
    dbpedia:Dave_Bronconnier ;
governing_body
    dbpedia:Calgary_City_Council ;
...

Altogether 9,100,000 RDF triples extracted from 754,000 infoboxes

Multi-Lingual Abstracts

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

• Short abstracts
  • English: 1,637,622
  • German: 246,791
  • French: 206,085
  • Dutch: 133,746
  • Polish: 118,874
  • Italian: 113,950
  • Spanish: 112,417
  • Japanese: 106,610
  • Portuguese: 104,842
  • Swedish: 100,267
  • Chinese: 54,991

Christian Bizer: Linked Data, DBpedia and D2R Server (26.06.2007)
The DBpedia Dataset

- 1,600,000 concepts
- including
  - 58,000 persons
  - 70,000 places
  - 35,000 music albums
  - 12,000 films
- described by 93 million triples
- using 8,141 different properties.

- 557,000 links to pictures
- 1,300,000 links to relevant external web pages
- 207,000 Wikipedia categories
- 75,000 YAGO categories

Accessing the DBpedia Dataset over the Web

1. SPARQL Endpoint
2. Linked Data Interface
3. RDF Dumps for Download
The DBpedia SPARQL Endpoint

- http://dbpedia.org/sparql
- can answer SPARQL queries like
  - Give me all Sitcoms from 1980 that are set in NYC?
  - All tennis players from Moscow?
  - All German musicians that were born in Berlin in the 19th century?
- Provides two extensions to SPARQL
  - free-text search within titles and abstracts
  - COUNT()
- hosted on a OpenLink Virtuoso server
DBpedia Use Cases

1. Improving Wikipedia Search
2. Royalty-Free Data Source for other Applications
3. Interlinking-Hub for the Emerging Web of Data

Improving Wikipedia Search

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>player</td>
<td>currentClub</td>
<td>club</td>
</tr>
<tr>
<td>player</td>
<td>clubnumber</td>
<td></td>
</tr>
<tr>
<td>player</td>
<td>countryYear</td>
<td></td>
</tr>
<tr>
<td>club</td>
<td>capacity</td>
<td></td>
</tr>
<tr>
<td>country</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants

Click on a column heading to sort results.
10 results found in 0.0s.

<table>
<thead>
<tr>
<th>Nr.</th>
<th>?player</th>
<th>?country</th>
<th>GDP nominal (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ciroinho</td>
<td>Brazil</td>
<td>80354 18756000</td>
</tr>
<tr>
<td>2</td>
<td>Gonzalo Ferro</td>
<td>Chile</td>
<td>62000 16432674</td>
</tr>
<tr>
<td>3</td>
<td>Lukas Podolski</td>
<td>Poland</td>
<td>69901 3653669</td>
</tr>
<tr>
<td>4</td>
<td>Mark Gonzalez</td>
<td>Liverpool</td>
<td>45662 47425000</td>
</tr>
<tr>
<td>5</td>
<td>Michael Thurf</td>
<td>Eintracht</td>
<td>52000 83438000</td>
</tr>
<tr>
<td>6</td>
<td>Ramón Morales</td>
<td>Chivas</td>
<td>72400 107784179</td>
</tr>
<tr>
<td>7</td>
<td>Robin van Persie</td>
<td>Arsenal</td>
<td>60432 16336346</td>
</tr>
<tr>
<td>8</td>
<td>Stefano Mauri</td>
<td>S.S. Lazio</td>
<td>82056 58751711</td>
</tr>
</tbody>
</table>
2.2 W3C SWEO Linking Open Data Project

- Community effort to
  - publish various open-license databases as RDF on the Web
  - interlink data items between different data sources

Project Participants

- Universities
  - Freie Universität Berlin (DE)
  - SIMILE, Massachusetts Institute of Technology (US)
  - KMi, Open University (UK)
  - DERI (IRE)
  - DB Group, University of Pennsylvania (US)
  - Universität Leipzig (DE)
  - L3S, Universität Hannover (DE)
  - University of London (UK)

- Companies
  - OpenLink (UK)
  - Talis (UK)
  - Zitgist (US)
  - MetaWeb (US)
  - Mondeca (FR)

- Outreach
  - Tim Berners-Lee (W3C, MIT)
  - Ivan Hermann (W3C)
Datasets on the Web

- Over one billion RDF triples served on the Web
- Around 120,000 RDF links between data sources

Example RDF Links

- RDF Links from DBpedia to other Data Sources

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

  `<http://dbpedia.org/resource/Tim_Berners-Lee> owl:sameAs

- RDF Links from FOAF profiles to DBpedia

  `<http://www.w3.org/People/Berners-Lee/card#i> owl:sameAs

  `<http://richard.cyganiak.de/foaf.rdf#cygri> foaf:topic_interest
Searching DBpedia together with Linked Data

Participating in the Linking Open Data Project

- **Wiki Page**
  - http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

- **Mailing List**
  - http://simile.mit.edu/mailman/listinfo/linking-open-data

- **Participating in the project**
  - Put your name on the Wiki page
  - Subscribe to the mailing list
  - Do something useful

Christian Bizer: Linked Data, DBpedia and D2R Server (26.06.2007)
2.3 D2RQ and D2R Server

- Most data is stored in relational databases today.
- We developed two artifacts for publishing relational databases on the Semantic Web
  - The D2RQ Mapping Language
  - D2R Server
**Data Model, Schema and Syntactic Heterogeneity**

<table>
<thead>
<tr>
<th>People</th>
<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>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Papers</th>
<th>ID</th>
<th>title</th>
<th>confID</th>
</tr>
</thead>
<tbody>
<tr>
<td>312</td>
<td>D2R Server</td>
<td>132</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rel_People_Papers</th>
<th>PersonID</th>
<th>PaperID</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>312</td>
<td></td>
</tr>
</tbody>
</table>

**Value Correspondences**

- D2RQ is a language for writing these correspondences down.
Example D2RQ Mapping

ClassMap People

PropertyBridge
foaf:name

PropertyBridge
foaf:mbox

PropertyBridge
dc:author

PropertyBridge
swc:conference

PropertyBridge
rdf:type

ClassMap Papers

PropertyBridge
dc:title

PropertyBridge
swc:conference

PropertyBridge
rdf:type

refersTo

Class Map

<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@@".

Christian Bizer: Linked Data, DBpedia and D2R Server (26.06.2007)
**Property Bridge**

<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>ID</th>
<th>name</th>
<th>email</th>
<th>ID</th>
<th>title</th>
<th>confID</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Chris</td>
<td><a href="mailto:chris@bizer.de">chris@bizer.de</a></td>
<td>312</td>
<td>D2R Server</td>
<td>132</td>
</tr>
</tbody>
</table>

Rel_People_Papers

<table>
<thead>
<tr>
<th>PersonID</th>
<th>PaperID</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>312</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";
### Other Features of the D2RQ Mapping Language

- Conditional mappings
- Value translation tables
- Integration of custom value transformation functions
- Different performance optimization hints

#### D2R Server

![D2R Server Diagram]

- RDF Browsers
- SPARQL Clients
- HTML Browsers
- RDF
- D2R Server
- HTML
- Non-RDF Database
- D2RQ Mapping File

*Christian Bizer: Linked Data, DBpedia and D2R Server (26.06.2007)*
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 and Standardization

- Open source project (LGPL)
- Around 2400 downloads (225 in April 2007)
- Public Data Sources using D2R Server
  - Roller, Weblog Server, Sun Microsystems
  - Images of Fruitfly Embryogenesis, Berkeley Drosophila Genome Project
  - DBtune, University of London
  - DBLP Bibliography, FU Berlin
  - Information Systems Group, FU Berlin
- OEM distribution as part of the TopBraid Ontology Editor
- W3C standardization starts next year
  (first meeting at MIT in September)
Thanks!

Comments?
Questions?

This talk is online at