DBpedia - Extracting structured data from Wikipedia

Christian Becker, Freie Universität Berlin
Hello

- **Name**
  - Christian Becker (beckr)

- **Job**
  - PhD Student at Freie Universität Berlin

- **Wikipedia Involvement**
  - OpenStreetMap integration

- **Semantic Web Projects**
  - DBpedia and DBpedia Mobile
  - Marbles Browser
  - BBC interlinking project
  - D2RQ and D2R Server
  - flickr™ wrappr
DBpedia

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

Contributors
- Freie Universität Berlin (Germany)
- Universität Leipzig (Germany)
- OpenLink Software (UK)
- Linking Open Data Community (W3C SWEO)
Extracting Structured Information from Wikipedia
The DBpedia Dataset

- Web-scale identifiers for 2.6 million “things”
  - 80 languages
  - including at least 213,000 persons
  - 328,000 places

- Altogether 274 million pieces of information (RDF triples)
  - 29 million triples extracted from infoboxes
  - 609,000 links to pictures
  - 3,150,000 links to relevant external web pages
  - 4,900,000 links to other LOD datasets

- Categorizations
  - DBpedia Ontology with 170 classes and 940 properties, based on the most frequently used classes
  - 415,000 Wikipedia categories
  - 75,000 YAGO categories

- Updated from Wikipedia dumps; Live Update soon
Example: RDF from an Infobox

<http://dbpedia.org/resource/Calgary>
  dbpedia:native_name "Calgary" ;
  dbpedia:altitude "1048" ;
  geo:lat "51.044998" ;
  geo:long "-114.057220" ;
  dbpedia:population_city "988193" ;
  dbpedia:population_metro "1079310" ;
  mayor_name
    dbpedia:Dave_Bronconnier ;
  governing_body
    dbpedia:Calgary_City_Council ;
  ...
The DBpedia Ontology

- Hand-made
- 350 most frequently used templates ↔ 170 ontology classes
- 2350 template properties ↔ 940 ontology properties
- Currently tracking 882,000 resources (articles)
- We are currently moving the mappings to meta:
  - http://meta.wikimedia.org/wiki/DBpedia/ontology
- Great for cross-language fusion
  - e.g. comparing inhabitants of en:London with fr:Paris
    - en:London – settlement infobox
    - fr:Paris – Communes de France
DBpedia Use cases

1. Improving Wikipedia Search
2. Improve Wikipedia Data quality through better template knowledge
3. Data Source for other Applications and Mashups
4. Nucleus for the Emerging Web of Data
DBpedia Search: Buenos Aires...

http://dbpedia.neofonie.de
Filter for films about Buenos Aires...

- Osvaldo Soriano
  - 1997 in Buenos Aires.

- Plan 4
  - Plan 4 is a Heavy metal band from Buenos Aires, Argentina that formed in 2004.

- Mad (band)
  - Mad is a hard rock band from Buenos Aires, Argentina. It was formed in 1997.
Refine: Films released between 2005 to 2008

<table>
<thead>
<tr>
<th>Your Filters</th>
<th>Reset Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>text search for buenos aires</td>
<td>item type Film</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Film</th>
<th>Year</th>
<th>Director</th>
</tr>
</thead>
<tbody>
<tr>
<td>La Chismosa</td>
<td>1938</td>
<td>Argentine film</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The film premiered in Buenos Aires.</td>
</tr>
</tbody>
</table>
3 Américas
3 Américas is a 2007 Argentine/US dramatic film written and directed by Cristina Kotz Cornejo. Produced by Cristina Kotz Cornejo and Angela Counts. The story takes place in Boston. Massachusetts and Buenos Aires, Argentina.

Buenos Aires 100 Kilómetros
Buenos Aires 100 kilómetros is an Argentine, French, and Spanish, film, written and directed by Pablo José Meza. The picture features Juan Ignacio Perez Roca, Emiliano Fernández, Alan Ardel, Hernan Wainstein, Juan Pablo Bazzini, among others. The film, partly funded by INCAA, tells a coming-of-age story in a town 100 kilometers from Buenos Aires.
Use DBpedia locations as starting points ...
Choose between 14 languages ...
Brandenburg Gate

Review of Brandenburg Gate, by fatorange
It’s a magnificent spot in Berlin that is associated with many important historic events. The city has recently put a lot of effort into its renovation.

Homepage: http://www.stadtentwicklung.berlin.de/denkmal/denk...

Based near of Christian Becker

The Brandenburg Gate is a former city gate and one of the main symbols of Berlin, Germany. It is located west of the city center at the intersection of Unter den Linden and Ebertstraße, immediately west of the Pariser Platz. It is the only remaining gate of a
Accessing the DBpedia dataset

- **Linked Data**
  - e.g. http://dbpedia.org/resource/London

- **SPARQL Interface**
  - http://dbpedia.org/sparql

- **Download RDF Dumps**

- **Amazon Public Datasets**
Linked Data Interface

- The project follows the Linked Data principles
  - All concepts are identified using URI references
  - All URIs are dereferencable over the Web into a small RDF snippet

- The Linked Data interface can be used by
  - Semantic Web Browsers, like
    - Tabulator Browser
    - Marbles
    - OpenLink RDF Browser
  - Semantic Web Crawlers, like
    - Zitgist (Zitgist LLC, USA)
    - SWSE (DERI, Ireland)
    - Swoogle (UMBC, USA)
<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<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="http://en.wikipedia.org/wiki/Berlin" alt="Image of Berlin" /></td>
<td>G2</td>
</tr>
<tr>
<td>is birthplace of</td>
<td>Adolf von Baeyer</td>
<td>G2</td>
</tr>
<tr>
<td>is birthplace of</td>
<td><a href="http://dbpedia.org/resource/person/Albert_Speer_%28the_younger%29">http://dbpedia.org/resource/person/Albert_Speer_%28the_younger%29</a></td>
<td>G2</td>
</tr>
</tbody>
</table>
The DBpedia SPARQL Endpoint

- http://dbpedia.org/sparql
- hosted on a OpenLink Virtuoso server
- can answer SPARQL queries like
  - Give me all Sitcoms that are set in NYC?
  - All German musicians that were born in Berlin in the 19th century?
  - All tennis players from Moscow?
  - All films by Quentin Tarentino?
  - All soccer players with tricot number 11, playing for a club having a stadium with over 40,000 seats and is born in a country with over 10 million inhabitants?
DBpedia and the web of data
The Web is a single information space build on open standards and hyperlinks.
Use RDF and HTTP to
1. publish structured data on the Web,
2. set links between data from one data source to data within other data sources.
What can I do with this?

Linked Data Browsers

Search Engines

Linked Data Mashups

HTTP

Thing

HTTP

Thing

data link

A

data link

B

data link

C

data link

D

data link

E
Example: Integrating Wikipedia content into BBC News

Based on simple tags, a BBC News article about Barack Obama in Berlin could automatically import Berlin photos and Weblinks for Barack Obama.
Mockup: Integrating BBC content into Wikipedia

Berlin
From Wikipedia, the free encyclopedia

This article is about the capital of Germany. For other uses, see Berlin (disambiguation).

Berlin is the capital city and one of sixteen states of Germany. With a population of 3.4 million within its city limits, Berlin is the country's largest city. It is the second most populous city and the ninth most populous urban area in the European Union.[2] Located in northeastern Germany, it is the centre of the Berlin-Brandenburg metropolitan area, comprising 5 million people from over 180 nations.[3]

First documented in the 13th century, Berlin was successively the capital of the Kingdom of Prussia (1701-1918), the German Empire (1871-1918), the Weimar Republic (1919-1933) and the Third Reich (1933-1945).[4] After World War II, the city was divided: East Berlin became the capital of East Germany while West Berlin became a Western enclave, surrounded by the Berlin Wall from 1961-1989.[5] Following German reunification in 1990, the city regained its status as the capital of all Germany.[6]

Berlin is a major center in European politics, culture, media, and science.[7][8] Its economy is primarily based on the service sector, encompassing a diverse range of creative industries, media corporations, environmental services, congress and convention venues; it also serves as a continental hub for air and rail transport.[9][10] Berlin is the third most visited tourist destination in the EU.[11] Other industries include traffic engineering, optoelectronics, IT, vehicle manufacturing, health care, biomedical engineering, and biotechnology.

The metropolis is home to world-renowned universities, research institutes, sporting events, orchestras, museums and personalities.[12] Berlin’s urban landscape and historical legacy has made it a popular setting for international film productions.[13] The city is recognized for its festivals, diverse architecture, nightlife, contemporary arts and a high quality of living.[14] Berlin has evolved into a global focal point for young individuals and artists attracted by a liberal lifestyle and modern zeitgeist.[15]
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

500 million RDF triples
120,000 data links between data sources

As of May 2007
LOD Datasets on the Web: September 2008
LOD Datasets on the Web: July 2009

4.5 billion triples
180 million data links
LOD Datasets on the Web: July 2009

6.2 billion triples
150 million data links
<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</td>
<td>76.000.000</td>
<td>1,1%</td>
<td>1.559.000</td>
<td>1,0%</td>
</tr>
<tr>
<td>Content</td>
<td>Cross-Domain</td>
<td>214.000.000</td>
<td>3,2%</td>
<td>3.992.000</td>
</tr>
</tbody>
</table>

| Total             | 6.726.000.000       | 148.948.000|

+ 2 billion triples from Data.gov published recently.
Linked Data Adoption

Who's publishing / supporting Linked Data? (excerpt)

- BBC
- Cyc Foundation
- Freebase
- MusicBrainz
- Thomson Reuters
- UK Ordnance Survey
- Zemanta
- Life Sciences Community
- Announced: New York Times and CNET

Yahoo! and Google have started to crawl Linked Data in its RDFa serialization as well as Microformats.
DBpedia interlinks (exemplary)

- MusicBrainz
- OpenStreetMap (via LinkedGeoData.org)
- BBC Programmes, Music
- Semantic CrunchBase
- Geonames
- WordNet
- World Factbook
- EuroStat
- Flickr (via flickrwrappr)
- FreeBase
- OpenCyc
- US Census
- Amazon, Google Base (via Book Mashup)
- Dailymed, Diseasome, Drugbank, Sider
Thanks

Questions?

References

- DBpedia.org
- LinkedData.org
- DBpedia Browser: http://dbpedia.neofonie.de
- http://meta.wikimedia.org/wiki/DBpedia/ontology
- Tim Berners-Lee’s TED Talk
- DBpedia Mobile:
  http://beckr.org/DBpediaMobile