# Wikimania 2009 Buenos Aires, August 26, 2009

# DBpedia - Extracting structured data from Wikipedia

Christian Becker, Freie Universität Berlin

### Hello

- Name
  - Christian Becker (beckr)
- Job
  - PhD Student at Freie Universität Berlin
- Wikipedia Involvment
  - 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

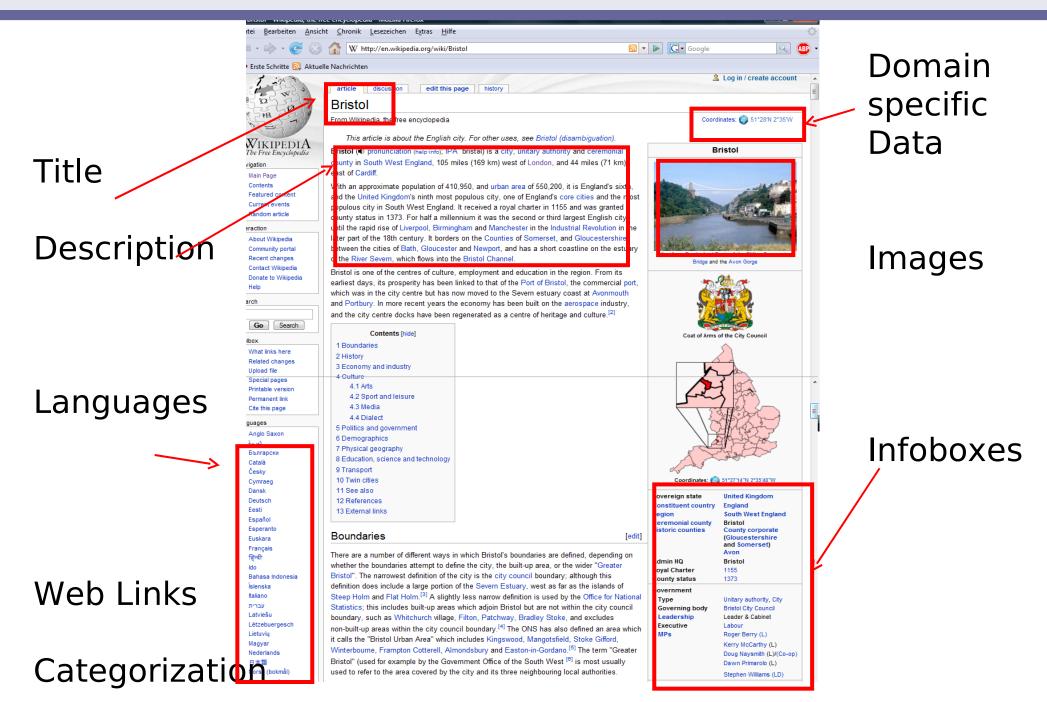
- Freie Universität Berlin (Germany)
- Universität Leipzig (Germany)
- OpenLink Software (UK)
- Linking Open Data Community (W3C SWEO)







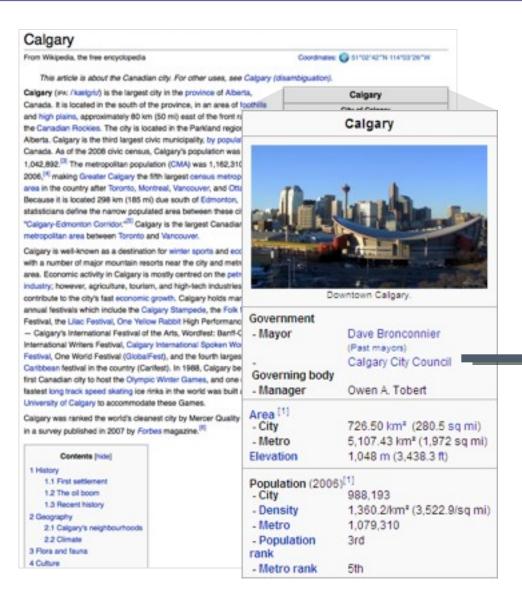
# **Extracting Structured Information from Wikipedia**



# **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...**





About DBpedia About neofonie

Contact

First | Previous | Next | Last

Results 1 to 1 of 1

### ▼ item type

start typing...

Place (227059)

Person (205390) Work (189671)

PopulatedPlace (161634)

Area (115930)

MusicalWork (102047)

Species (90789)

Organisation (76089)

Athlete (75791)

Album (74089)

Artist (69050)

FootballPlayer (50947)

City (40190)

Film (34683)

MusicalArtist (34483)

Actor (25923)

Single (24612)

Building (23071)

EducationalInstitution (21052)

Company (20173)

OfficeHolder (19974)

Book (15679)

RadioStation (13032)



The Free Encyclopedia

**Your Filters** 

No Filters selected

### Wikipedia

Wikipedia is a free, multilingual encyclopedia project operated by the non-profit Wikimedia Foundation. Its name is a portmanteau of the words wiki and encyclopedia. Wikipedia's 10 million articles, about a quarter of which are in English, have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone who can access the Wikipedia website. Launched in 2001 by Jimmy Wales and Larry Sanger, it is currently the largest and most popular general reference work on the Internet. Critics of Wikipedia target its systemic bias and inconsistencies and its policy of favoring consensus over credentials in its editorial process. Wikipedia's reliability and accuracy are also an issue. Other criticisms are centered on its susceptibility to vandalism and the addition of spurious or unverified information.

# http://dbpedia.neofonie.de

## Filter for films about Buenos Aires...



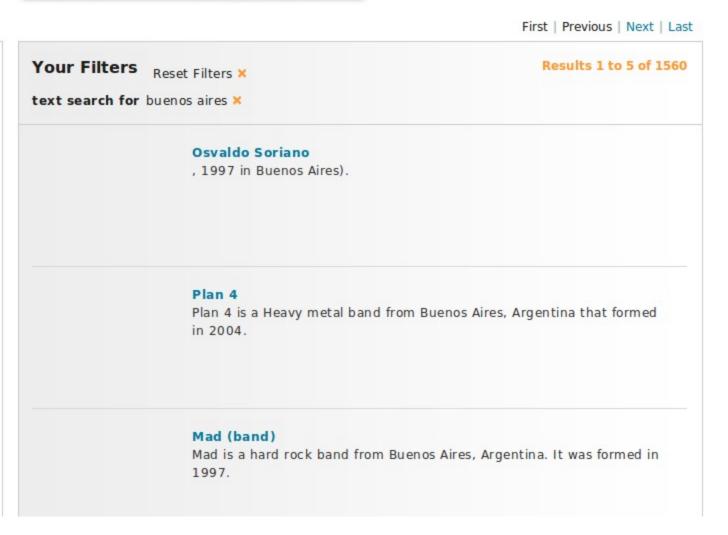
enter search terms... Search

About DBpedia About neofonie

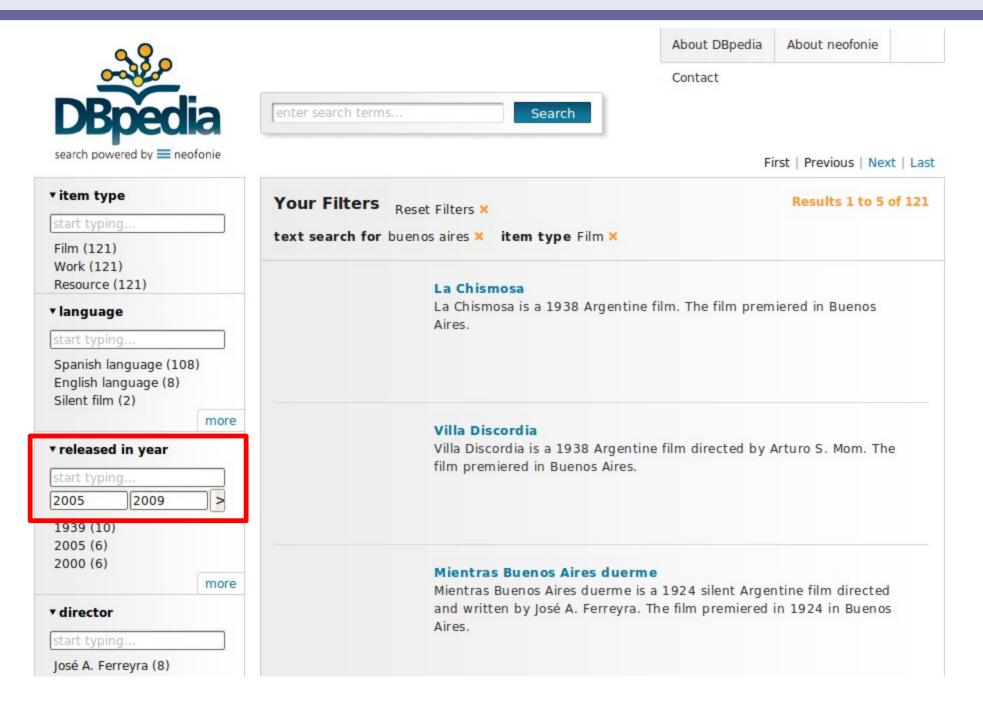
Contact

▼ item type start typing... Resource (1560) Person (918) Athlete (565) FootballPlayer (470) Place (259) PopulatedPlace (222) Area (206) Artist (201) Work (181) Organisation (166) Film (121) Sports leam (103) SoccerClub (103) Actor (91) OfficeHolder (80) MusicalArtist (72) RugbyPlayer (37) MusicalWork (35) Building (33) Album (32) Company (28) EducationalInstitution (23)

FormulaOneRacer (23)



## Refine: Films released between 2005 to 2008



# **Results link to Wikipedia**



enter search terms... Search

Reset Filters X

About DBpedia About neofonie

Contact

First | Previous | Next | Last

Results 1 to 5 of 8

### ▼ item type start typing... Film (8) Work (8) Resource (8) ▼ language start typing.. Spanish language (6) English language (2) ▼ released in year start typing... 2005 (6) 2006 (1) 2007 (1) ▼ director

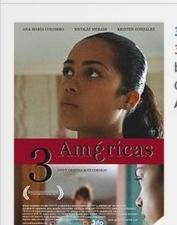
start typing...

Jorge Gaggero (1)

Pablo José Meza (1)

Israel Adrián Caetano (2)

more



**Your Filters** 

### 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**

text search for buenos aires X item type Film X released in year 2005 to 2009 X

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 ...**



## **Get more details ...**



# **Accessing the DBpedia dataset**

- Linked Data
  - e.g. http://dbpedia.org/resource/London
- SPARQL Interface
  - http://dbpedia.org/sparql
- Download RDF Dumps
  - http://wiki.dbpedia.org/Downloads
- 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)

## Berlin

URI: http://dbpedia.org/resource/city/Berlin

Go!

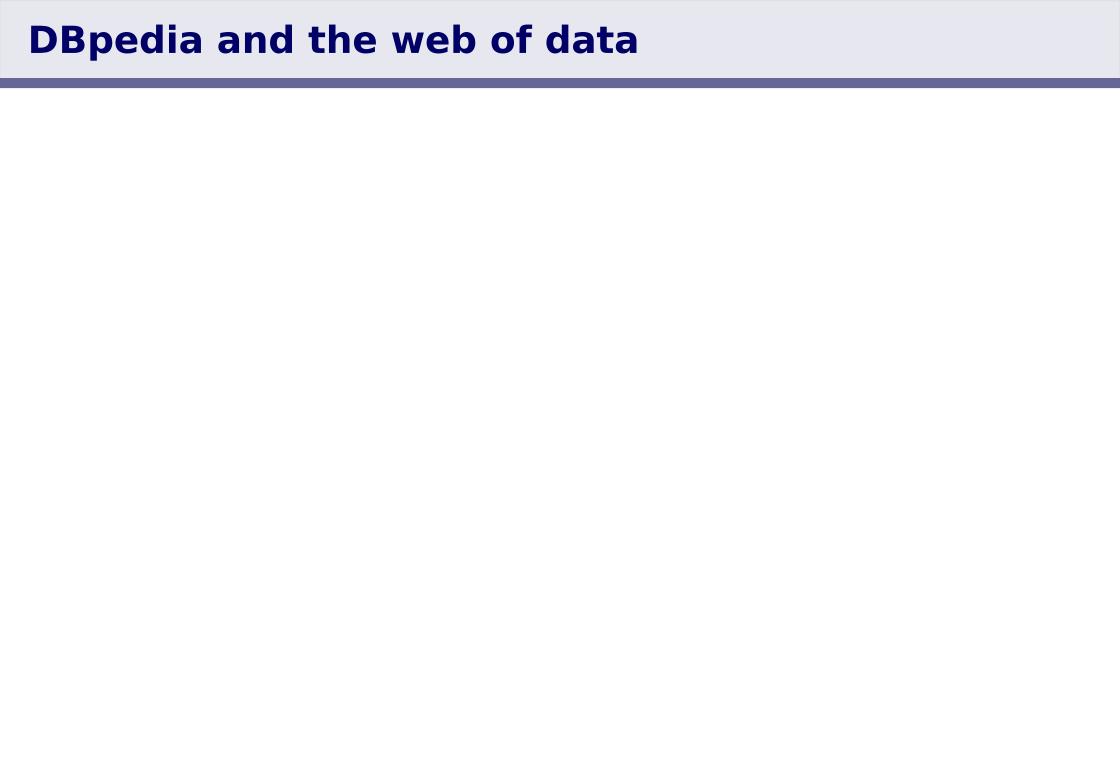
Property	Value	Sources		
population	3398888	<u>G2</u>		
type	http://dbpedia.org/City@	<u>G2</u>		
comment	Berlin is the capital city and one of the sixteen Federal States of Germany. It is the country's largest city in area and population, and the second most populous city in the European Union.			
comment	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.			
label	Berlin			
sameAs	http://sws.geonames.org/2950159/ ₪			
subject	http://dbpedia.org/resource/category/Berlin ❷			
subject	http://dbpedia.org/resource/category/Capitals_in_Europe ₽			
subject	http://dbpedia.org/resource/category/Cities_in_Germany ❷			
subject	http://dbpedia.org/resource/category/German_state_capitals@			
subject	http://dbpedia.org/resource/category/Host_cities_of_the_Summer_Olympic_Games @			
subject	http://dbpedia.org/resource/category/States_of_Germany.ø			
sourceURL	Berlin 🗗	<u>G1</u>		
depiction		<u>G2</u>		
page	http://en.wikipedia.org/wiki/Berlin ₽	<u>G2</u>		
is birthplace of	Adolf von Baeyer ❷	<u>G2</u>		
is birthplace of	http://dbpedia.org/resource/person/Albert_Speer_%28the_younger%29@	G2		





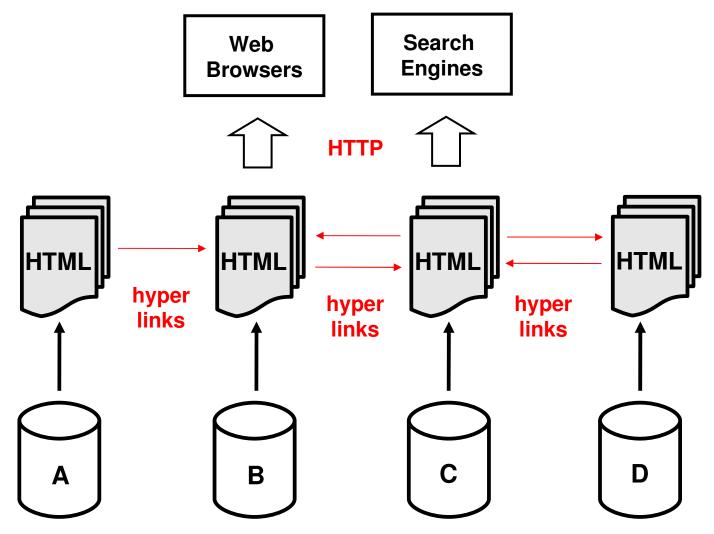
# The DBpedia SPARQL Endpoint

- 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?



### The Web of Documents

# The Web is a single information space build on open standards and hyperlinks.

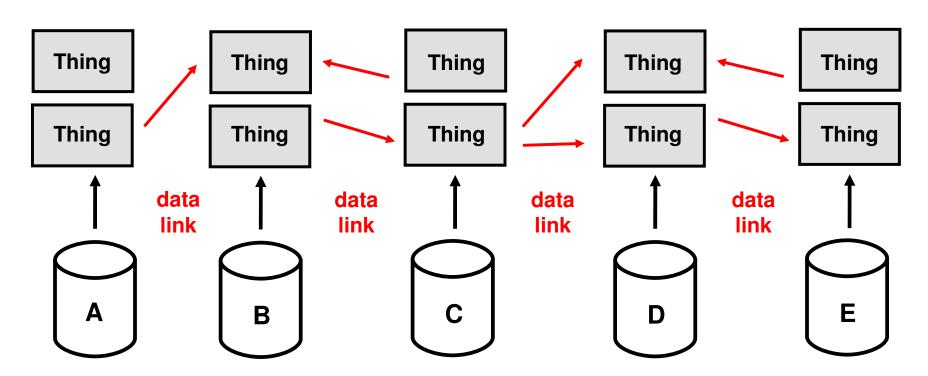


Christian Becker: DBpedia - Extracting structured data from Wikipedia (Buenos Aires, 08/26/2009)

### **Linked Data**

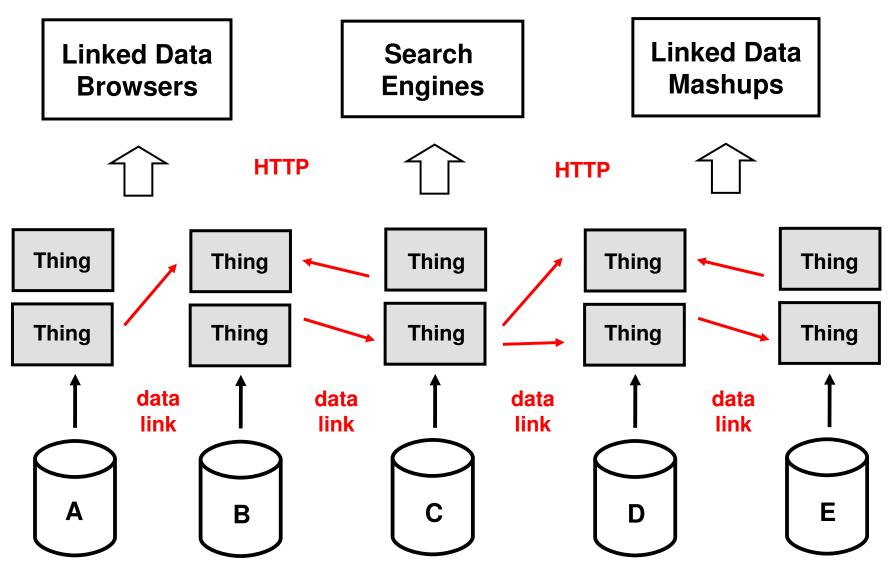
### **Use RDF and HTTP to**

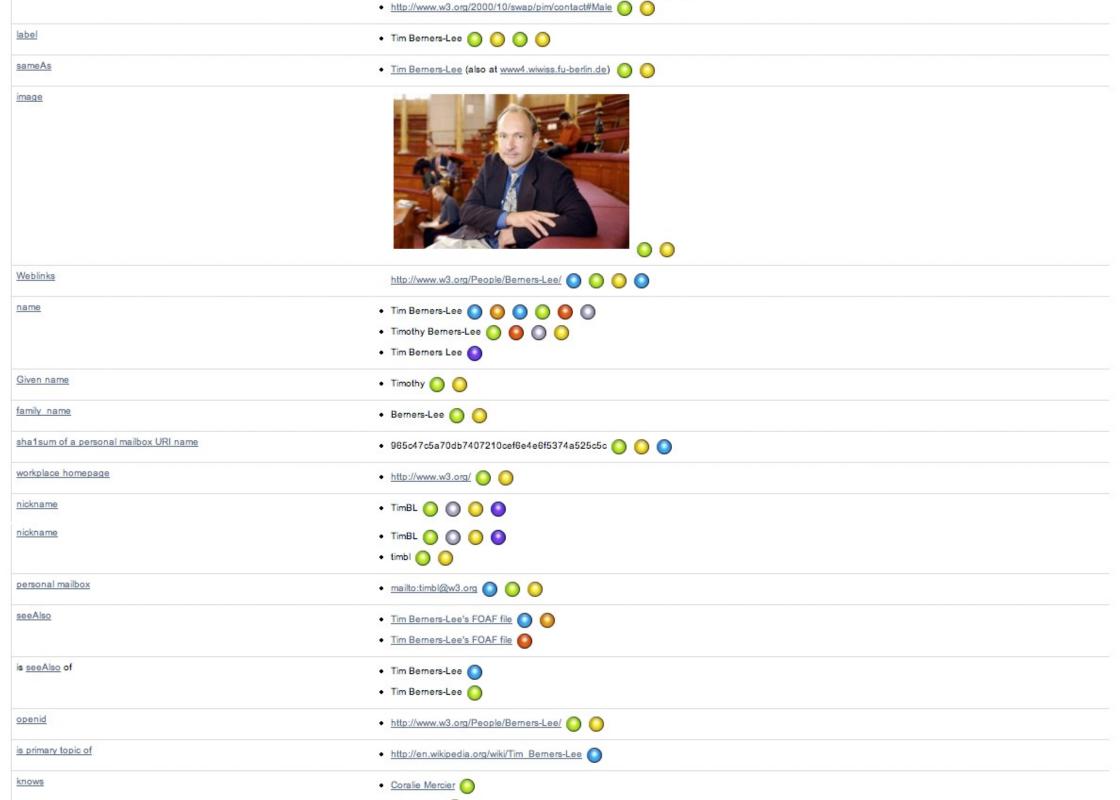
- 1. publish structured data on the Web,
- 2. set links between data from one data source to data within other data sources.



## What can I do with this?







## **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

WIKIPEDIA
The Free Encyclopedia

### navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

### search



### interaction

- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia
- Donate to Wikipedia
- Help

### toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this page

### languages

- \* Afrikaans
- Alemannisch

Help us provide free content to the world by donating today!

article discussion edit this page

history

Log in / create account

### 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]

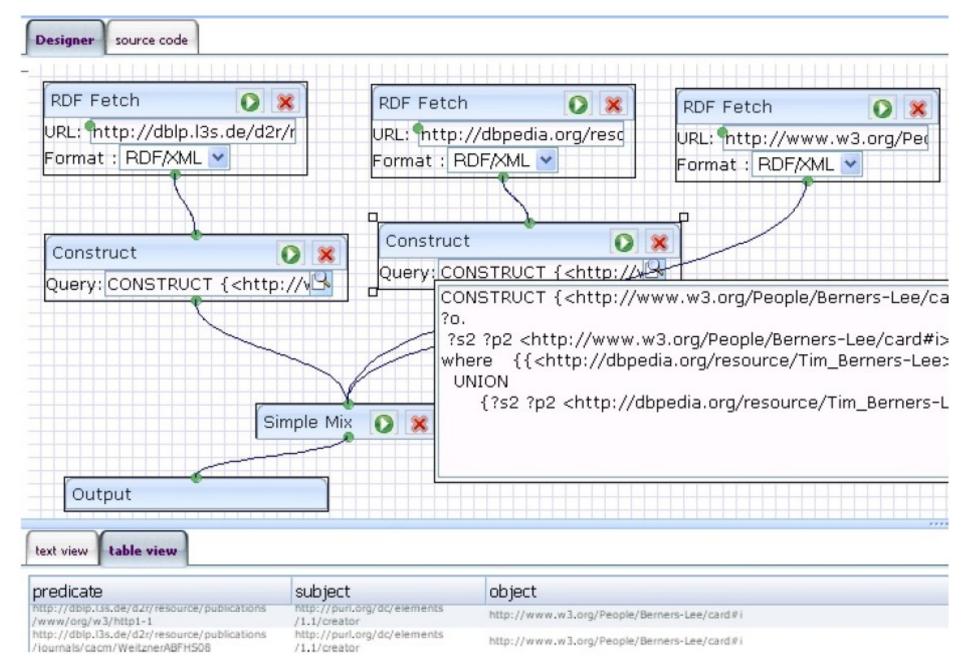
Berlin is a major center in European politics, culture, media, and science. [7][8] Its economy is primarily based on the service sector, encompassing a diverse range of creative industries, media corporations, environmental services, congress and convention venues; it also serves as a continental hub 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]

# Contents [hide] 1 History 1.1 17th to 19th century



## **DERI Semantic Web Pipes**

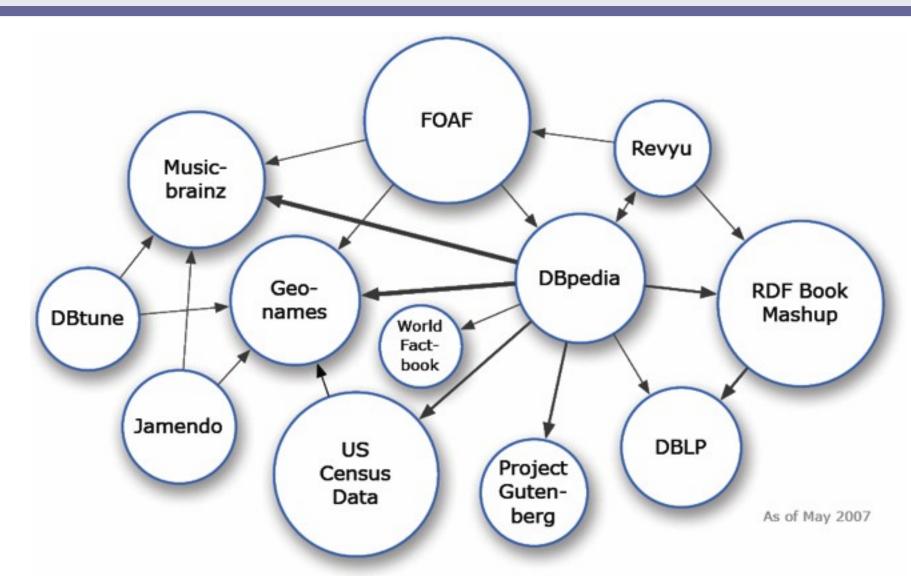


# **W3C Linking Open Data Project**



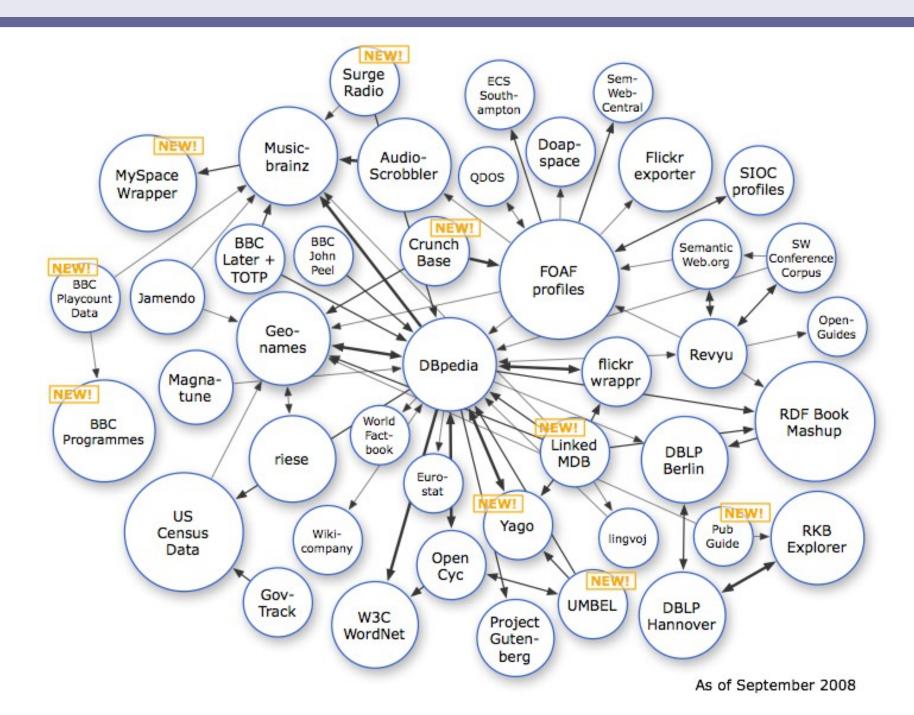
- Community effort to
  - publish existing open license datasets as Linked Data on the Web
  - interlink things between different data sources

# **LOD Datasets on the Web: May 2007**

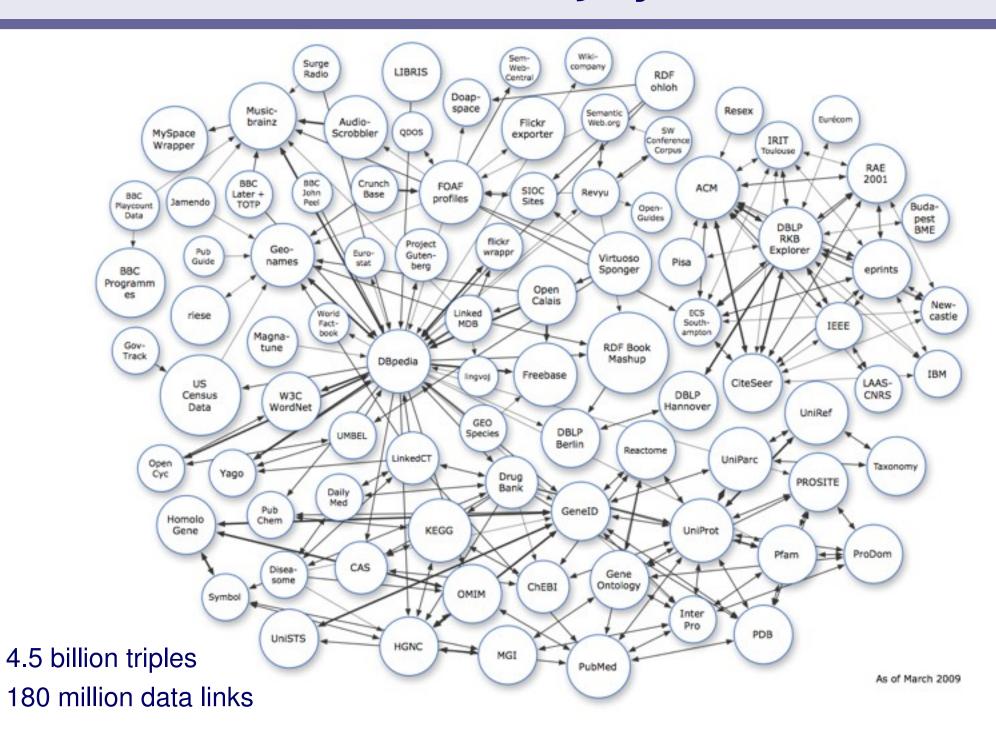


500 million RDF triples120,000 data links between data sources

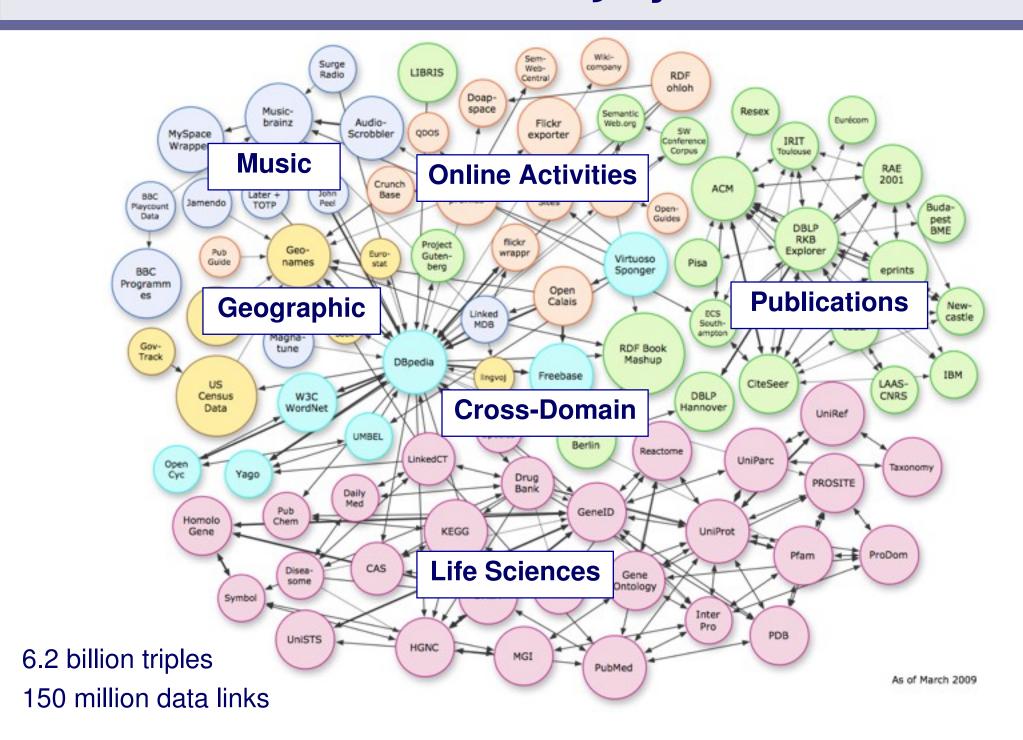
# **LOD Datasets on the Web: September 2008**



# **LOD Datasets on the Web: July 2009**



# **LOD Datasets on the Web: July 2009**



# LOD data set statistics as of July 2009

Domain	No of Triples	% of	No of Links	% of
		Cloud		Links
Media	698.000.000	10,4%	1.238.000	0,8%
Publications	212.000.000	3,2%	4.922.000	3,3%
Life Sciences	2.429.000.000	36,1%	133.199.000	89,4%
Geographic Data	3.097.000.000	46,0%	4.038.000	2,7%
User Generate Content	76.000.000	1,1%	1.559.000	1,0%
Cross-Domain	214.000.000	3,2%	3.992.000	2,7%
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)
- Dailvmed, 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
     http://www.ted.com/index.php/talks/tim\_berners\_lee\_on\_the\_next\_web.html
  - DBpedia Mobile: http://beckr.org/DBpediaMobile

