16th International World Wide Web Conference
Developers Track, May 11, 2007

DBpedia

Querying Wikipedia like a Database

Chris Bizer, Freie Universität Berlin
Sören Auer , Universität Leipzig
Georgi Kobilarov, Freie Universität Berlin
Jens Lehmann, Universität Leipzig
Richard Cyganiak, Freie Universität Berlin

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 datasets on the Web
- **Contributors**
  - Freie Universität Berlin (Germany)
  - Universität Leipzig (Germany)
  - OpenLink Software (UK)
  - Linking Open Data Community (W3C SWEO)

Outline

1. Extracting Structured Information from Wikipedia
2. The DBpedia Dataset
3. Accessing the DBpedia Dataset over the Web
4. Use Cases
   1. Improving Wikipedia Search
   2. Royalty-Free Data Source for other Applications
   3. Interlinking-Hub for the Emerging Web of Data

Extracting Structured Information from Wikipedia

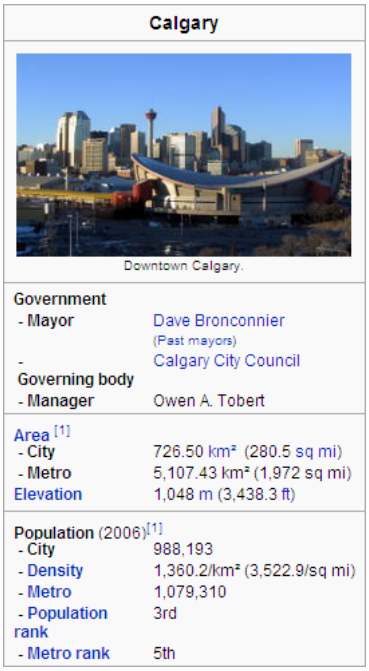
- **Wikipedia consists of**
  - 6.9 million articles
  - in 251 languages
  - monthly growth-rate: 4%
- **Wikipedia articles contain structured information**
  - infoboxes which use a template mechanism
  - images depicting the article's topic
  - categorization of the article
  - links to external webpages
  - 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



Extracting Other Article Data

- Short and long abstracts in 10 different languages

```
dbpedia:Calgary dbpedia:abstract “Calgary is the largest ...”@en ;
                dbpedia:abstract “Calgary ist eine Stadt ...”@de .
```

- Categorization information

```
dbpedia:Calgary skos:subject dbpedia:Category_Cities_in_Alberta ;
                skos:subject dbpedia:Host_cities_Olympic_Games .
```

- Links to the original Wikipedia articles, pictures and relevant external web pages

```
dbpedia:Calgary
      foaf:page <http://en.wikipedia.org/wiki/Calgary> ;
      dbpedia:wikipage-de <http://de.wikipedia.org/wiki/Calgary> ;
      foaf:depiction <http://upload.wikimedia.org/thumb/3/32> ;
      dbpedia:reference <http://www.calgary.ca> ;
      dbpedia:reference <http://www.tourismcalgary.com>.
```

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

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

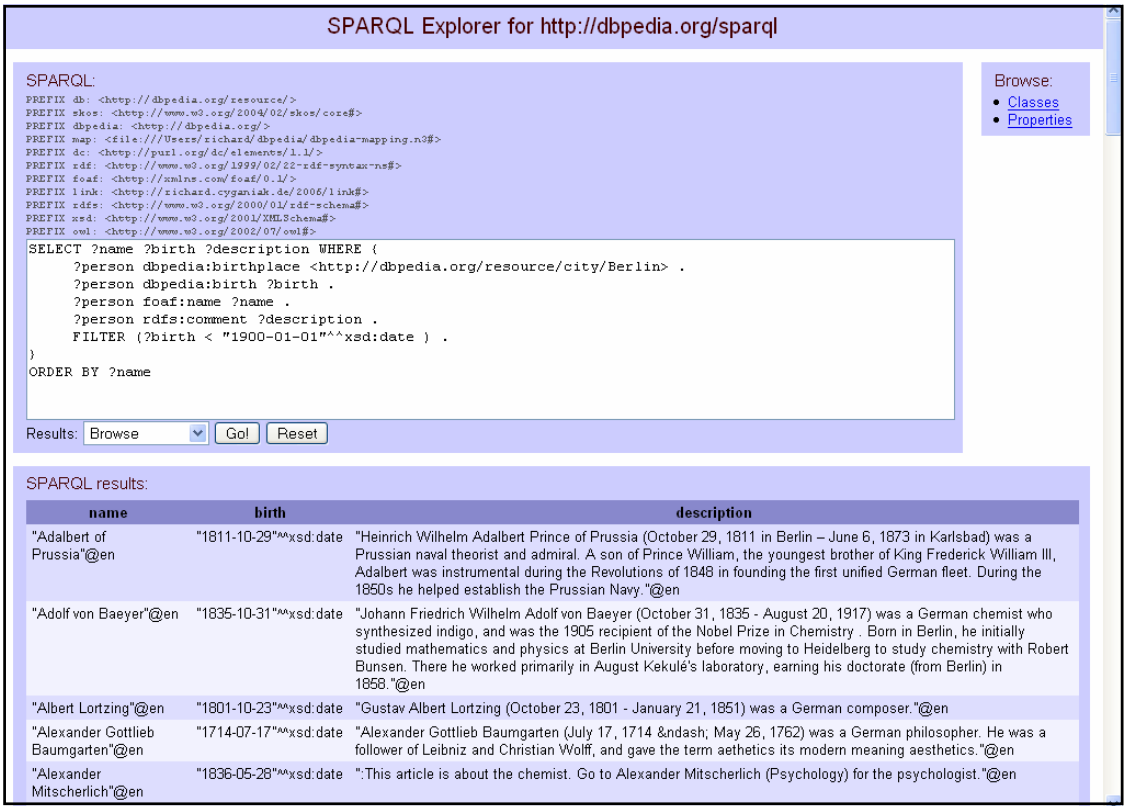
Property	Value
abstract	{{zh-tw:卡加利;zh-cn:卡尔加里}} {{Alzh-tw:卡加利;zh-cn:卡尔加里}}(Calgary)是一座位于加拿大亞伯達省南部落基山脚的城市。面积789.0平方公里，海拔约1048米，人口约100万，是艾伯塔省经济、金融、和文化中心。卡尔加里一词的意思是"清澈流动的水"。十八世纪七十年代，开始有欧洲探险者在此定居，后来成为西北骑警警(North West Mounted Police)的一间营地，后加拿大太平洋铁路横贯其北，卡尔加里成军发展成市。1947年这里发现了丰富的石油和天然气，从此城市面积了迅速的发展。世界上众多的包括中国的石油公司都在这里设有常驻机构，很多大的石油公司的总部就或设立在这里，因此卡尔加里也被称作加拿大的能源中心。1988年在这里举办过第十五届冬季奥林匹克运动会。当时市为奥林的室内滑冰场位于知名学府卡尔加里大学的校区内。这个城市的工程师圆屋顶全加第一，多次被评为世界上最干净的城市。七月间，一年一度的牛仔节(Stampede)是这个城市最著名的节日，每年都能吸引来自世界各地总主要是美国的游客。城市以西的班夫国家公园(Banff National Park)是世界著名的自然风景区，也是加拿大第一共国家公园，包括加拿大落基山脉。卡尔加里大都会地区是加拿大最好才城市之一。卡尔加里年中国的大步皇帝城 城市。"
abstract	Calgary (gälisch für Klares Wasser) ist eine Stadt in Kanada in der Provinz Alberta. Sie ist die drittgrößte und am schnellsten wachsende Großstadt Kanadas und hat 956.078 Einwohner (Stand 2005)."
abstract	Calgary es la mayor ciudad de la provincia de Alberta, Canadá. Se ubica en una región de colinas y altiplanicies, a aproximadamente 80 km al este de las Montañas Rocosas. Tercera ciudad de Canadá en término de población, contaba según el censo de abril de 2006 con 991.759 habitantes. La población estimada de su área metropolitana era de 1.060.300 habitantes en 2005 (véase Región de Calgary), lo que la convierte en la quinta mayor de Canadá. El "corredor Calgary-Edmonton" es la región urbana más poblada situada entre Toronto y Vancúver. Su nombre proviene del de una playa situada en la isla de Mull, en Escocia. Los habitantes de Calgary se llaman, en inglés, "Calgarians". La ciudad de Calgary es un destino muy conocido para los deportes de invierno y el ecoturismo: cerca de la ciudad se sitúa una gran cantidad de importantes lugares de vacaciones. La economía de Calgary se centra sobre todo en la industria petrolífera, aunque la agricultura, el turismo y la alta tecnología también contribuyen al rápido desarrollo económico de la ciudad. Calgary es la anfitriona de varios festivales anuales, como la Calgary Stampede, el Folk Music Festival, el Lilac Festival, el GlobalFest y el segundo festival de cultura canbeña en importancia del país, el Carifest. En 1988, Calgary se convirtió en la primera ciudad canadiense en acoger los Juegos Olímpicos de Invierno."
abstract	Calgary est la plus grande ville de la province de l'Alberta (Canada). Elle se situe dans le sud de la province, dans une région de collines et de plateaux à environ 80 km à l'est des montagnes Rocheuses. Troisième ville du Canada, en terme de population, elle comptait selon le recensement d'Avril 2006, 991 759 habitants (1 000 300 avec son agglomération en 2005, ce qui la met au cinquième rang des agglomérations canadiennes). Le « corridor Calgary-Edmonton » est la région urbaine la plus peuplée entre Toronto et Vancouver. Elle tire son nom d'une plage située sur l'île de Mull en Ecosse. Les habitants de Calgary sont appelés « Calgariens ». La ville de Calgary est une destination bien connue pour les sports d'hiver et l'écotourisme ; un grand nombre de stations de vacances importantes se situent près de la ville. L'économie de Calgary est surtout centrée sur l'industrie pétrolière ; toutefois, l'agriculture, le tourisme et la haute technologie contribuent également au développement économique rapide de la ville. Calgary est également l'hôte de plusieurs festivals annuels majeurs, dont le Stampede de Calgary, le Folk Music Festival, le Lilac Festival, le GlobalFest, et le deuxième festival de culture des Caraïbes en importance au pays (Carifest). En 1988, Calgary devient la première ville canadienne à accueillir les Jeux olympiques d'hiver. Calgary est la ville la plus prospère dans la province la plus riche du Canada. Anciennement, les barons étaient appelés les Cheiks aux yeux bleus."
abstract	Calgary is de grootste stad van de Canadese provincie Alberta, en ligt op ongeveer 80 kilometer afstand van de Canadese Rocky Mountains. Calgary had in april 2006 een inwonental van 991.759 en in de agglomeratie wonen 1.060.300 mensen. Het is de op drie na grootste stad van Canada, na Toronto, Montréal en Vancouver. De inwoners van Calgary staan bekend als Calgarians. De stad is een populaire bestemming voor wintersportvakanties door het groot aantal gelegen nabijgelegen 'vakantieparken' in de bergen. De economie van de stad bestaat vooral uit de petroleumindustrie, ook al worden de landbouw, het toerisme en technologie steeds belangrijker. In 1988 organiseerde Calgary de Olympische Winterspelen. De burgemeester van de stad is Dave Bronconnier. Aan de "University of Calgary" studeren meer dan 28.000 studenten."
abstract	Calgary är en stad belägen i den södra delen av provinsen Alberta i Kanada. Calgary är Albertas folkrikaste stad med över 991,759 invånare, och Calgary är även den tredje största staden, invånarmässigt, i Canada. En invånare i Calgary kallas för Calgarian. Calgary är en viktigt resmål för vintersport och ekoturism med ett flertal stora berg och vildmarksområden i närheten av staden."
abstract	Calgary è una città che si trova nella provincia dell'Alberta, in Canada. È situata nel sud della regione, in una zona di colline e alte pianure, a est delle Montagne Rocciose (Rocky Mountains). Con i suoi 991.759 abitanti (stima del 2006) è la più grande città dell'Alberta e la quinta più grande in tutto il Canada."
abstract	Calgary é a cidade mais populosa da província canadense de Alberta, e a terceira mais populosa do país, sendo que sua região metropolitana é a quinta mais populosa do país. Localiza-se no sul da província, a cerca de 80 quilômetros leste das Montanhas Rochosas. Sua população é de 991 759 habitantes, possuindo aproximadamente 1,06 milhão de habitantes na sua região metropolitana. Fundada em 1875, Calgary é atualmente um centro financeiro e comercial, onde estão localizadas as sedes das principais empresas petrolíferas do Canadá."
abstract	Calgary – największe miasto kanadyjskiej prowincji Alberta, leżące na przedgórzu Gór Skalistych nad rzeką Bow."
abstract	カルガリー（英： ）は、カナダ、アルバータ州の都市。アルバータ州最大の都市。1914年に石炭でカナダで最初の石油が発見されてから急速に発展した。石油経済では カナダの中心的役割を担い、大手石油会社のオフィスビルが建ち並ぶ。カナディアン・ロッキー観光の玄関口。1988年にはカルガリー冬季オリンピックの舞台となった。毎年7月のはじめには、カウボーイの祭典「カルガリー・スタンピード」が開催される。"

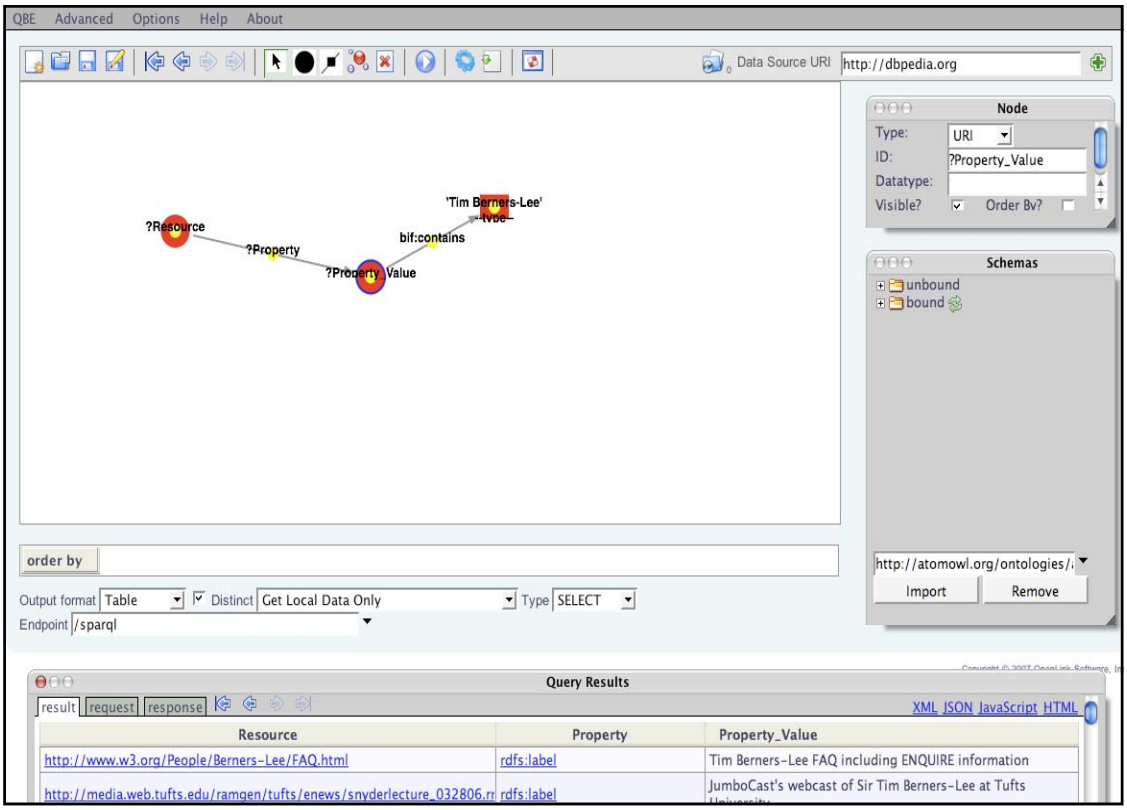
Accessing the DBpedia Dataset over the Web

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

The DBpedia SPARQL Endpoint

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


SPARQL Explorer for http://dbpedia.org/sparql
SPARQL:
PREFIX db: <http://dbpedia.org/resource/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX map: <file:///Users/richard/dbpedia/dbpedia-mapping.n3#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX link: <http://richard.cyganiak.de/2006/link#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT ?name ?birth ?description WHERE {
?person dbpedia:birthplace <http://dbpedia.org/resource/city/Berlin> .
?person dbpedia:birth ?birth .
?person foaf:name ?name .
?person rdfs:comment ?description .
FILTER (?birth < "1900-01-01"^^xsd:date) .
}
ORDER BY ?name
Results: Browse
Go!
Reset
Browse:
Classes
Properties
SPARQL results:
name
birth
description
"Adalbert of Prussia"@en
"1811-10-29"^^xsd:date
"Heinrich Wilhelm Adalbert Prince of Prussia (October 29, 1811 in Berlin – June 6, 1873 in Karlsbad) was a Prussian naval theorist and admiral. A son of Prince William, the youngest brother of King Frederick William III, Adalbert was instrumental during the Revolutions of 1848 in founding the first unified German fleet. During the 1850s he helped establish the Prussian Navy."@en
"Adolf von Baeyer"@en
"1835-10-31"^^xsd:date
"Johann Friedrich Wilhelm Adolf von Baeyer (October 31, 1835 - August 20, 1917) was a German chemist who synthesized indigo, and was the 1905 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 1858."@en
"Albert Lortzing"@en
"1801-10-23"^^xsd:date
"Gustav Albert Lortzing (October 23, 1801 - January 21, 1851) was a German composer."@en
"Alexander Gottlieb Baumgarten"@en
"1714-07-17"^^xsd:date
"Alexander Gottlieb Baumgarten (July 17, 1714 – May 26, 1762) was a German philosopher. He was a follower of Leibniz and Christian Wolff, and gave the term aethetics its modern meaning aesthetics."@en
"Alexander Mitscherlich"@en
"1836-05-28"^^xsd:date
":This article is about the chemist. Go to Alexander Mitscherlich (Psychology) for the psychologist."@en

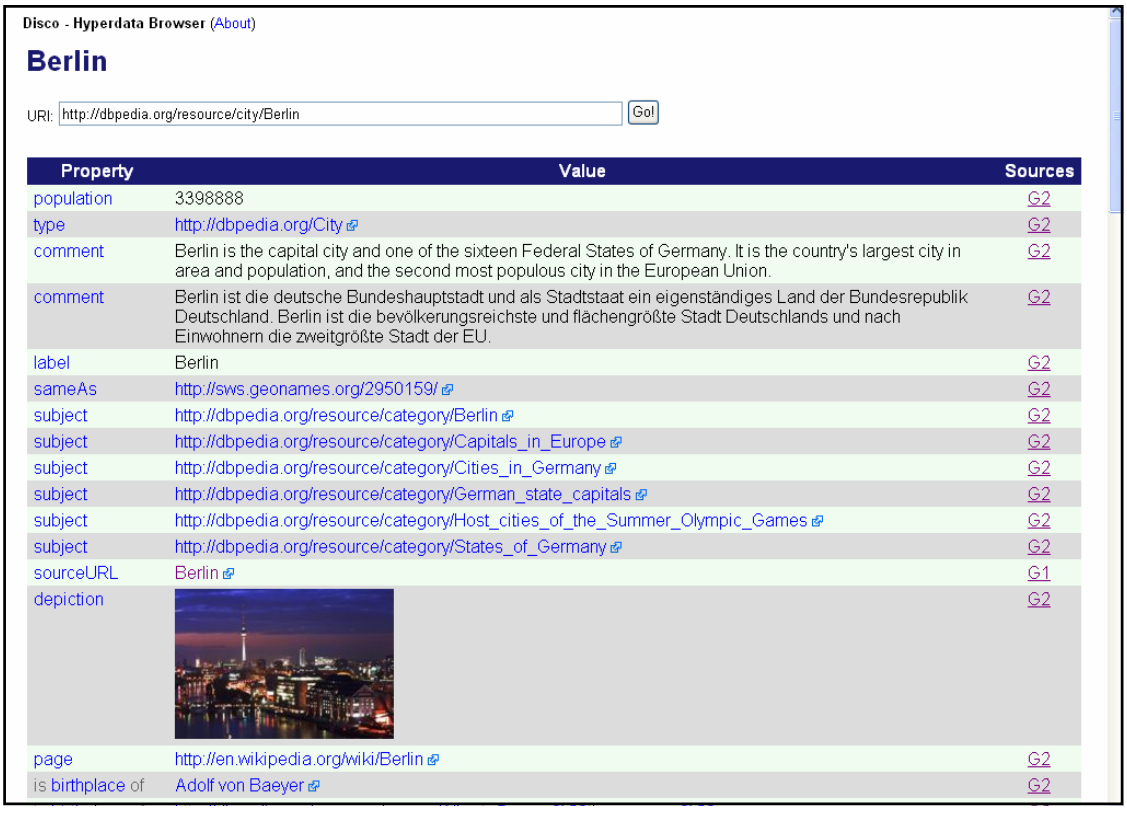

QBE
Advanced
Options
Help
About
Data Source URI
http://dbpedia.org
?Resource
?Property
?Property_Value
bif:contains
'Tim Berners-Lee'
Node
Type:
URI
ID:
?Property_Value
Datatype:
Visible?
Order By?
Schemas
unbound
bound
order by
Output format
Table
Distinct
Get Local Data Only
Type
SELECT
Endpoint
/sparql
http://atomowl.org/ontologies/
Import
Remove
Query Results
result
request
response
XML
JSON
JavaScript
HTML
Resource
Property
Property_Value
http://www.w3.org/People/Berners-Lee/FAQ.html
rdfs:label
Tim Berners-Lee FAQ including ENQUIRE information
JumboCast's webcast of Sir Tim Berners-Lee at Tufts

The 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
  - This snippets contain all in-bound and out-bound arcs of a node

- **The Linked Data interface can be used by**
  - Semantic Web Browsers, like
    - DISCO Hyperdata Browser
    - Tabulator Browser
    - OpenLink RDF Browser
  - Semantic Web Crawlers, like
    - SWSE (DERI, Ireland)
    - Zitgist (Zitgist LLC, USA)
    - Swoogle (UMBC, USA)

Christian Bizer et al: DBpedia – Querying Wikipedia Like a Database (May 11, 2007)



DBpedia Use Cases

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

Christian Bizer et al: DBpedia – Querying Wikipedia Like a Database (May 11, 2007)

Improving Wikipedia Search

Improving Wikipedia Search

UNIVERSITÄT LEIPZIG pedia

Query Wikipedia

This semantic database contains over 10 million statements extracted from the English Wikipedia.

search for queries | Most popular | Upcoming

- Tennis players from Moscow
- Sitcoms set in NYC
- Soccer player with tricot nr. 11, playing for a club having a stadium with >40.000 seats, born in a country with >10M inhabitants
- People influenced by Friedrich Nietzsche
- Films longer than 5 hours
- Space Missions
- Film music composer born 1965
- People being 1.80m tall
- List of Web browser software
- Mayors of US cities higher than 1000m
- Pictures of American guitarists
- Battles in Saxony
- What connects Innsbruck and Leipzig
- Hip hop CDs from Texas Artists
- Scientists and their doctoral advisors

<< 1 >>

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

Subject	Predicate	Object	
?player	currentclub	?club	[-]
?player	clubnumber	11	[-]
?player	countryofbirth	?country	[-]
?club	capacity	>40000	[-]
?country		>10000000	[-]

[+]

GDP_PPP (11)
population_estimate (10)
population_census (9)
established_date2 (8)
established_date1 (6)
established_date3 (5)
GDP_nominal (3)
accessionEUdate (2)

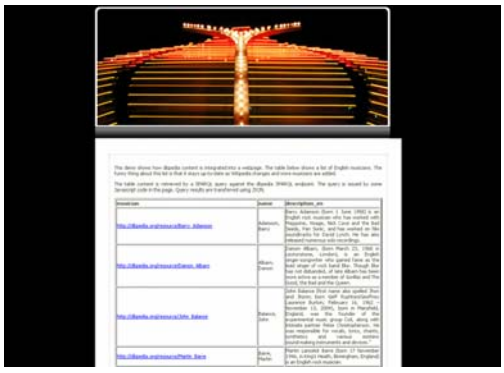
Click on a column hea… this page. Results: 10

10 results found in 0.0…

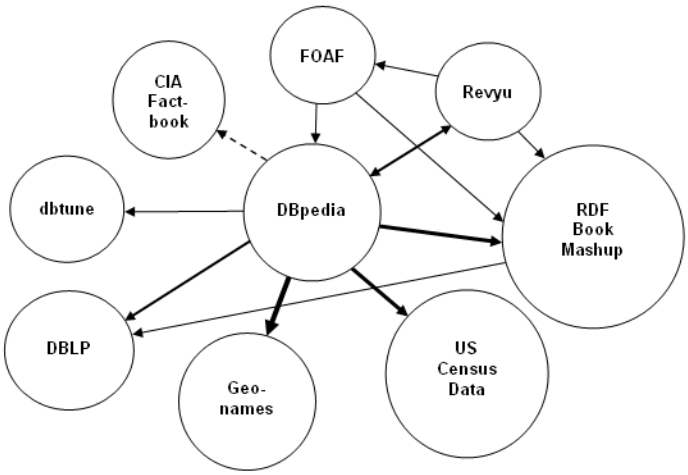
Nr.	?player		?country	>40000	>10000000
1	Cicinho	Real Madrid	Brazil	80354	187560000
2	Gonzalo Fierro	Colo-Colo	Chile	62000	16432674
3	Lukas Podolski	FC Bayern Munich	Poland	69901	38536869
4	Mark González	Liverpool F.C.	South Africa	45362	47432000
5	Michael Thurk	Eintracht Frankfurt	Germany	52000	82438000
6	Ramón Morales	Chivas de Guadalajara	Mexico	72480	107784179
7	Robin van Persie	Arsenal F.C.	Netherlands	60432	16336346
8	Stefano Mauri	S.S. Lazio	Italy	82656	58751711

Royalty-Free Data Source for other Applications

- **DBpedia is published under GNU Free Documentation License**
- **Example use case: SPARQL generated tables within a webpage that update automatically**



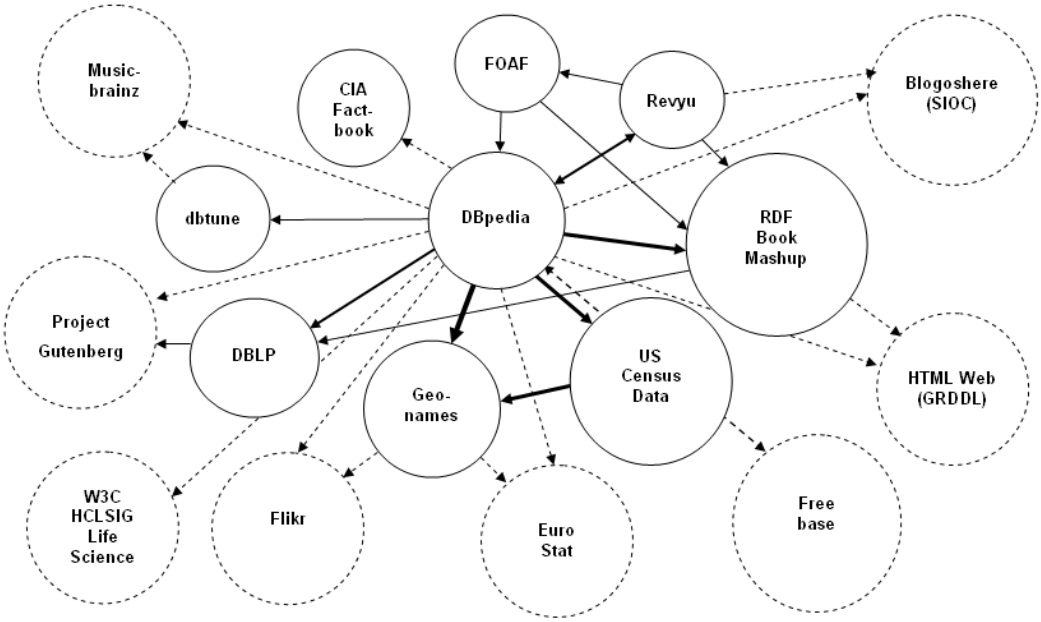
Interlinking-Hub for the Emerging Web of Data



- W3C SWEO Linking Open Data Project
- Over 600 million triples served as Linked Data
- Around 100,000 out-bound RDF links within DBpedia

Christian Bizer et al: DBpedia – Querying Wikipedia Like a Database (May 11, 2007)

Where will the project be in 2 month?



Christian Bizer et al: DBpedia – Querying Wikipedia Like a Database (May 11, 2007)

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
  - Talk to Kingsley Idehen, Tom Heath or me here at WWW2007
  - Publish and interlink open data

Example RDF Links

- **Out-Bound RDF Link**

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

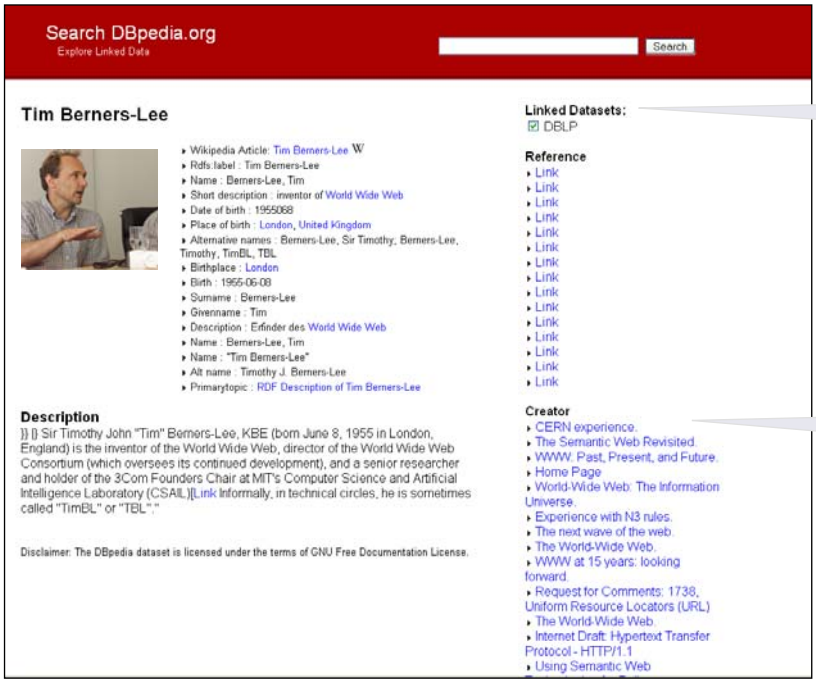
```
<http://dbpedia.org/resource/Tim_Berners-Lee> owl:sameAs
<http://www4.wiwiss.fu-berlin.de/dblp/resource/person/100007> .
```

- **In-Bound RDF Links**

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

```
<http://richard.cyganiak.de/foaf.rdf#cygri> foaf:topic_interest
<http://dbpedia.org/resource/Semantic_Web> .
```

Browsing DBpedia together with Linked Data



Linked Datasets

Linked Data from DBLP

Future Work

1. **Do a lot of data cleansing**
2. **Automate the data extraction process**
3. **Try to improve the classification**
4. **Interlink DBpedia with more datasets**

- **How can I support DBpedia?**
  - **Develop another cool user interface to DBpedia**
  - **Publish more RDF datasets with dereferencable URIs**
  - **Interlink your datasets with DBpedia**
  - **Talk to Kingsley Idehen, Tom Heath or me here at WWW2007 about participating in the Linking Open Data project**

Thanks!

DBpedia Website

http://dbpedia.org/docs/

Linking Open Data Project Website

http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData