Profiling Linked Open Data

Data Profiling and Data Cleansing

Anja Jentzsch
Information Systems Group (Prof. Dr. Felix Naumann)

Outline

- Introduction to Linked Data
  - □ Data Model
  - □ Data Variety
  - □ Example Data Set: DBpedia
- Profiling Linked Data
  - □ Challenges
  - □ Comparison: Traditional vs Linked Data Profiling
  - □ Existing Approaches

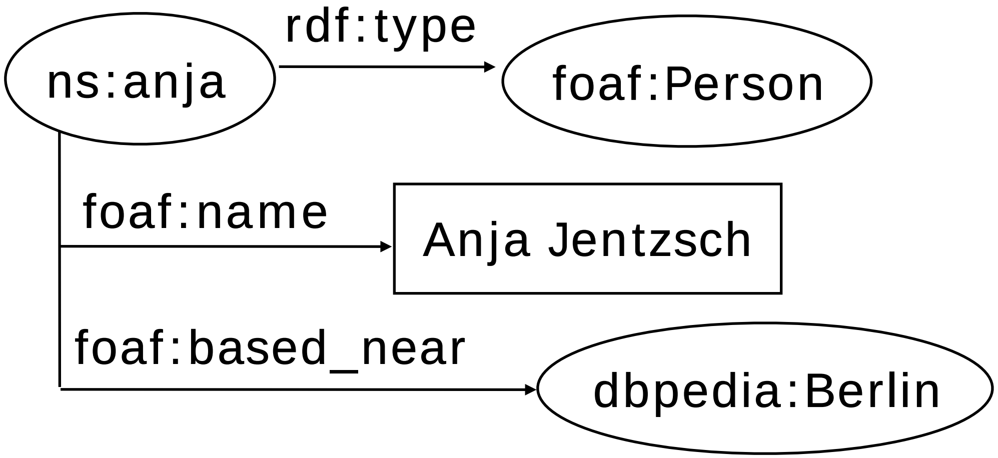
Linked Data Principles

Set of best practices for publishing structured data on the Web in accordance with the general architecture of the Web.

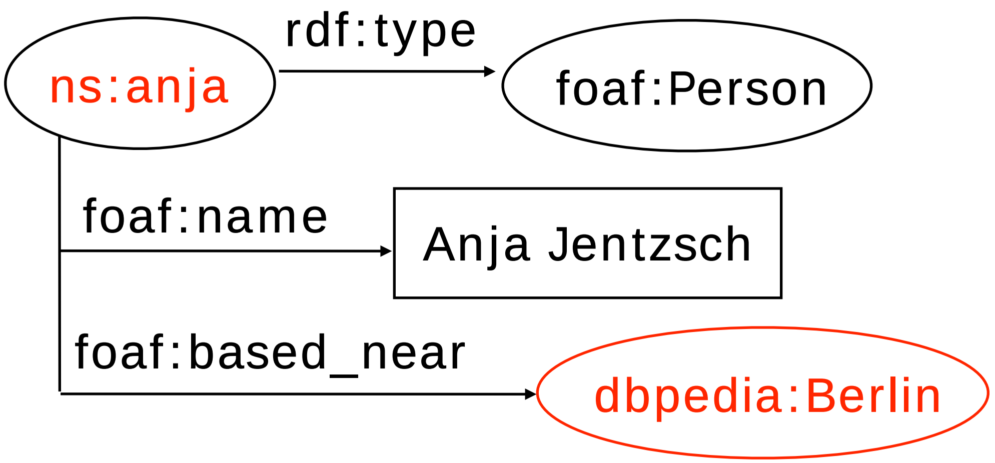
1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful RDF information.
4. Include RDF statements that link to other URIs so that they can discover related things.

Tim Berners-Lee, http://www.w3.org/DesignIssues/LinkedData.html, 2006

The RDF Data Model



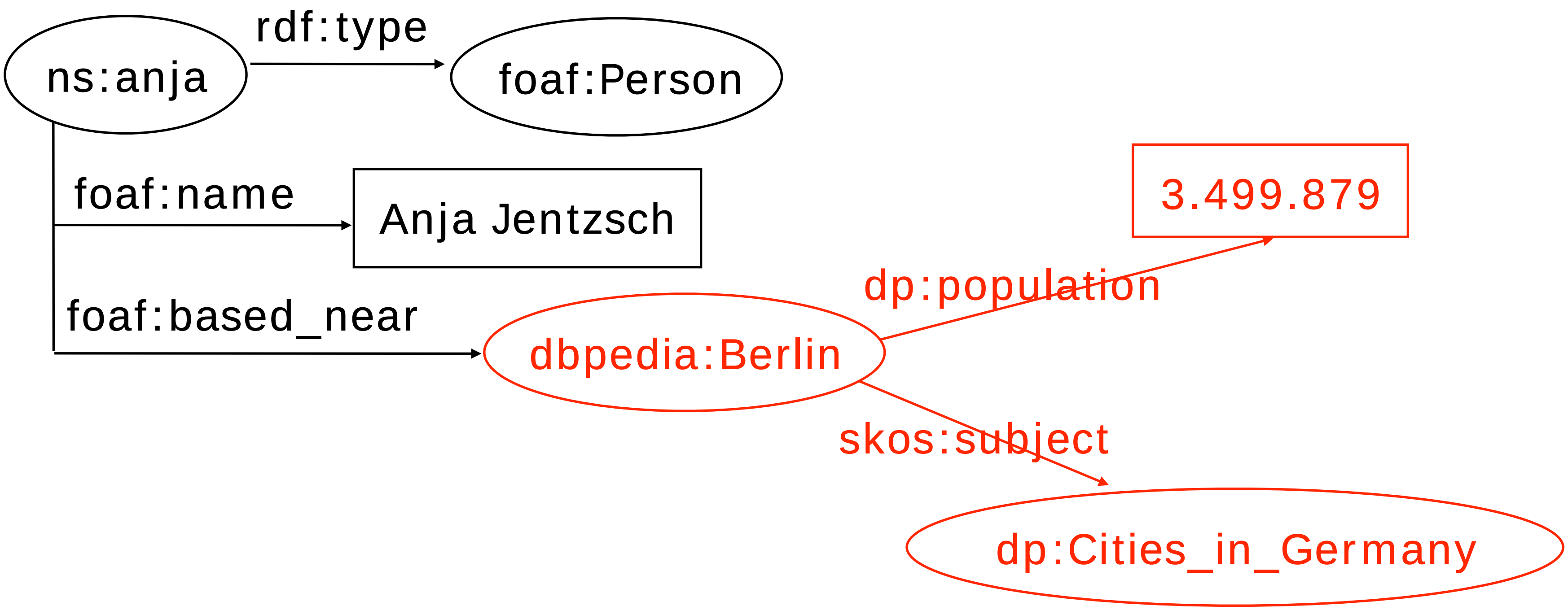
Identifying Data Items using URIs



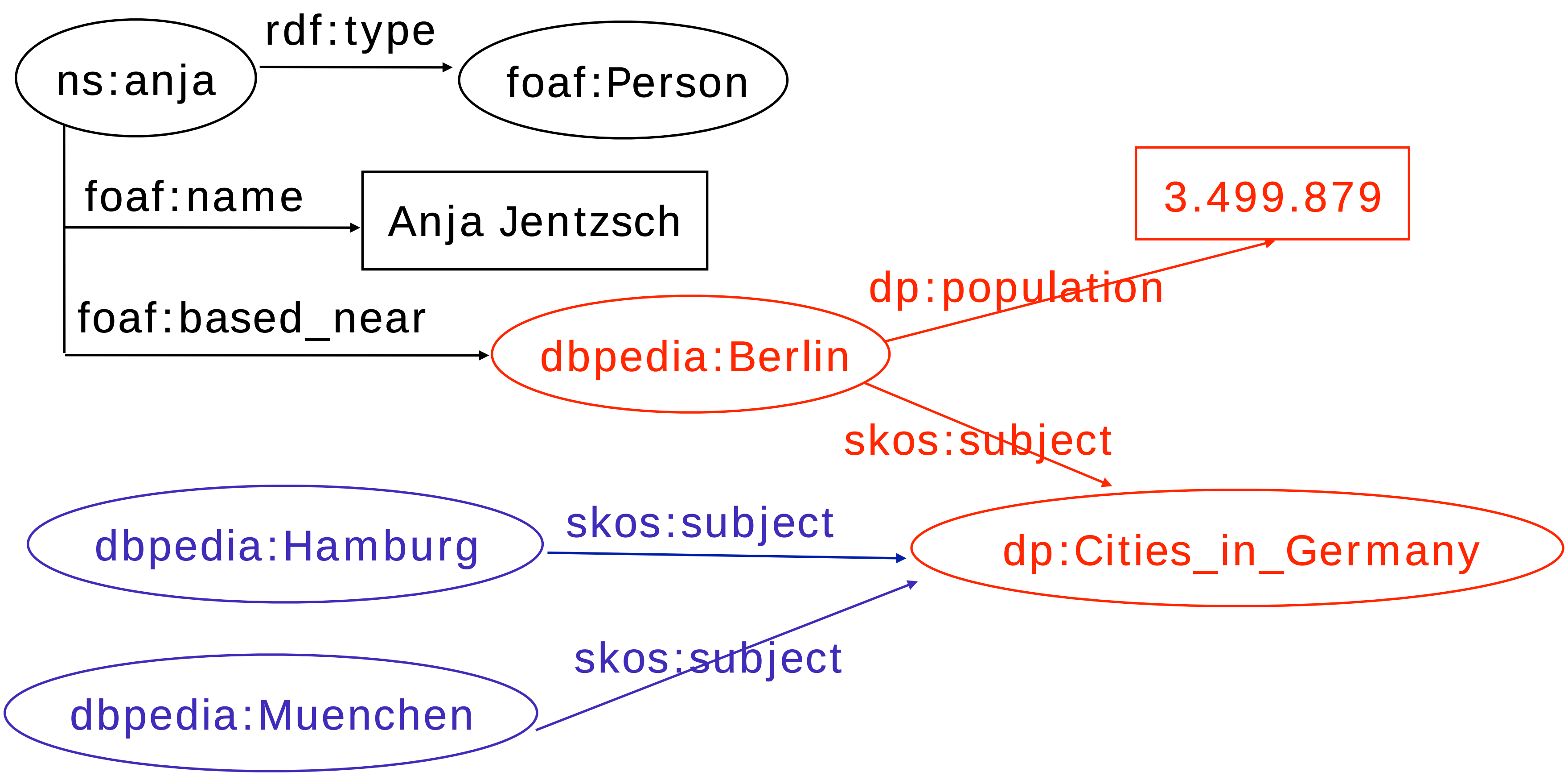
ns:anja = http://www.anjeve.de#anja

dbpedia:Berlin = http://dbpedia.org/resource/Berlin

Dereferencing URIs over the Web



Dereferencing URIs over the Web



RDF Representation Formats

- RDF/XML

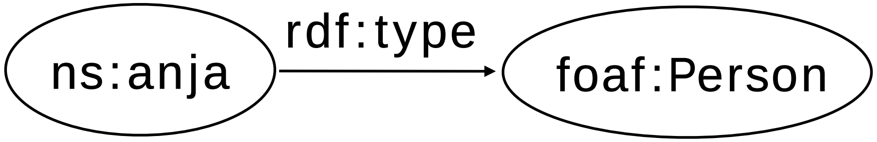
```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
   xmlns:foaf="http://xmlns.com/foaf/0.1/">
<foaf:Person rdf:about="http://anjeve.de#anja">
    <foaf:name>Anja Jentzsch</foaf:name>
</foaf:Person>
```

- RDF N-Triples

```
<http://anjeve.de#anja> <http://www.w3.org/1999/02/22-
rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/
Person>  .
<http://anjeve.de#anja> <http://xmlns.com/foaf/0.1/
name> „Anja Jentzsch“ .
```

- RDF N3

RDF Representation Formats



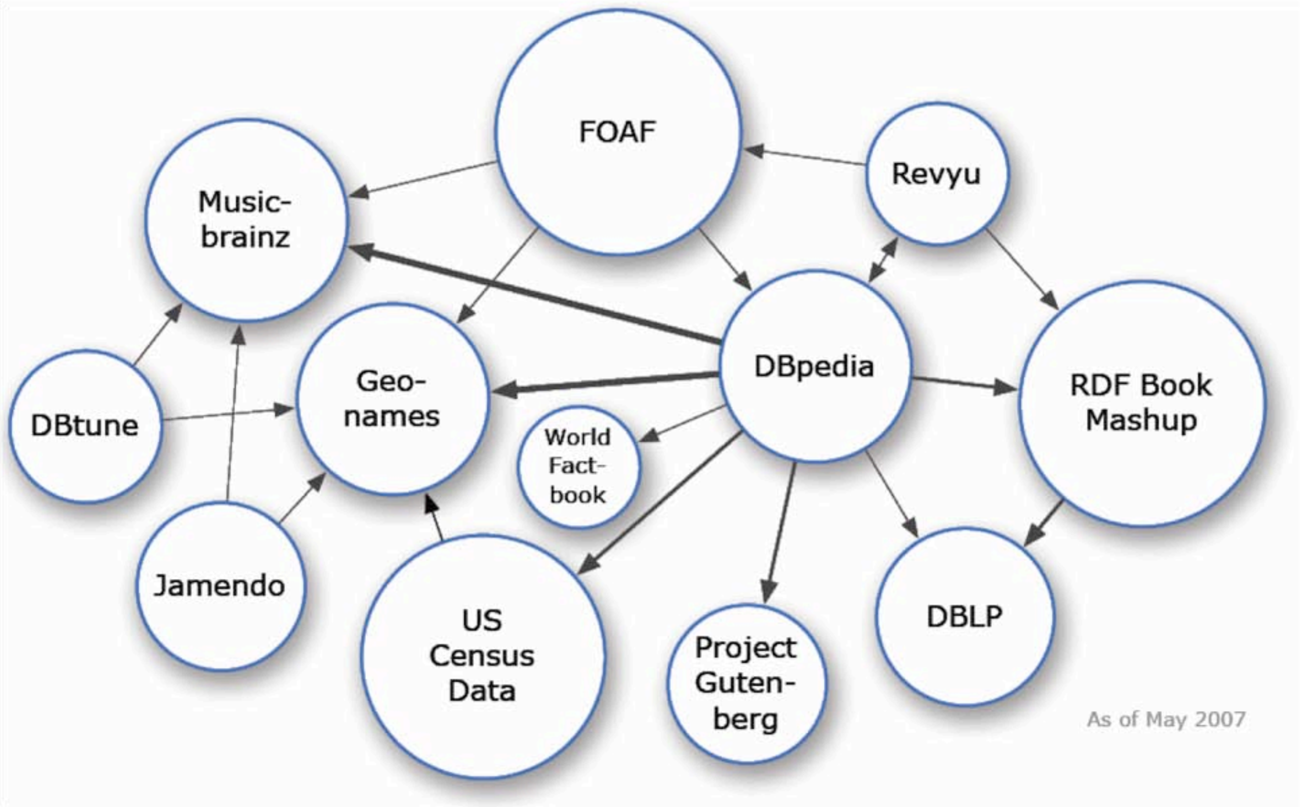
```
<http://anjeve.de#anja> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://xmlns.com/foaf/0.1/Person>  .
```

- <Subject> <Predicate> <Object>
- In the end it’s all triples!

Properties of the Web of Linked Data

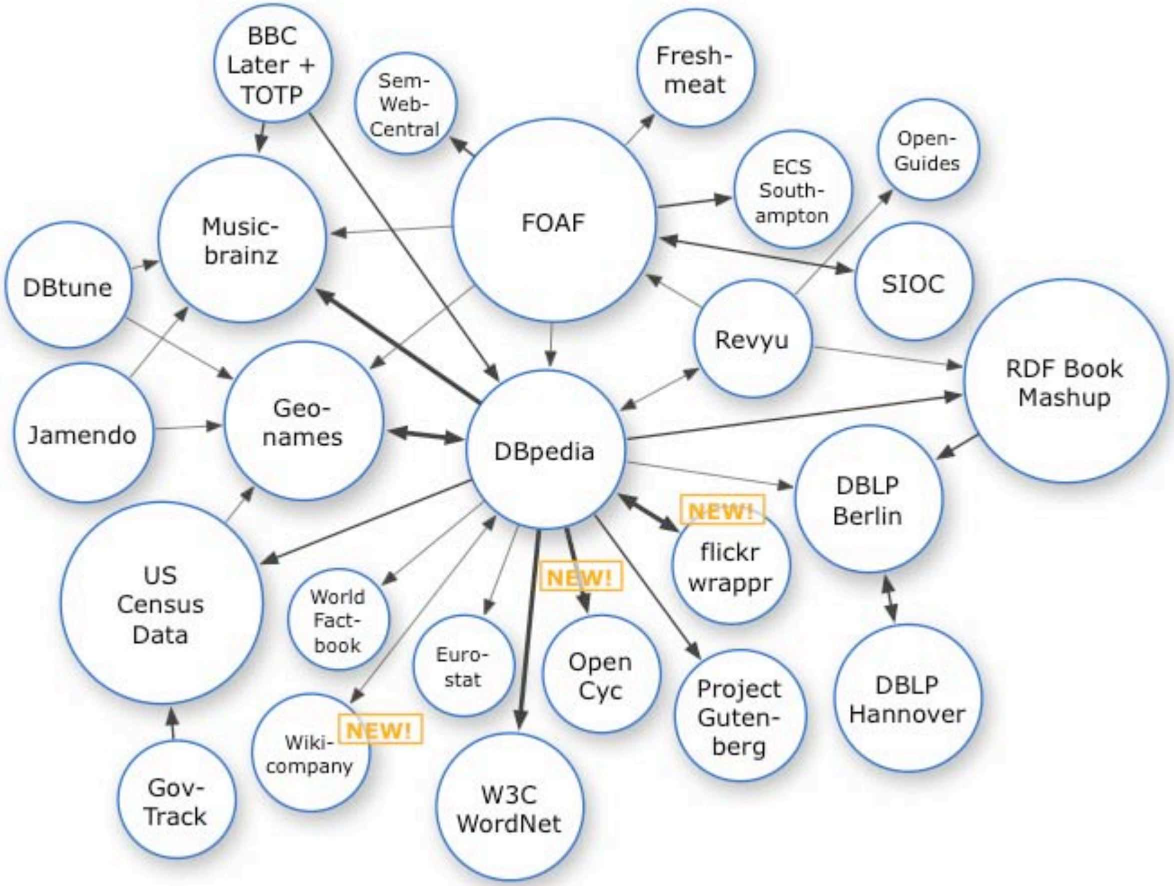
- Global, distributed dataspace build on a simple set of standards
  - RDF, URIs, HTTP
- Entities are connected by links
  - Creating a global data graph that spans data sources and
  - Enables the discovery of new data sources
- Provides for data-coexistence
  - Everyone can publish data to the Web of Linked Data
  - Everyone can express their personal view on things
  - Everybody can use the vocabularies/schemas that they like

Web of Data (as of May 2007)



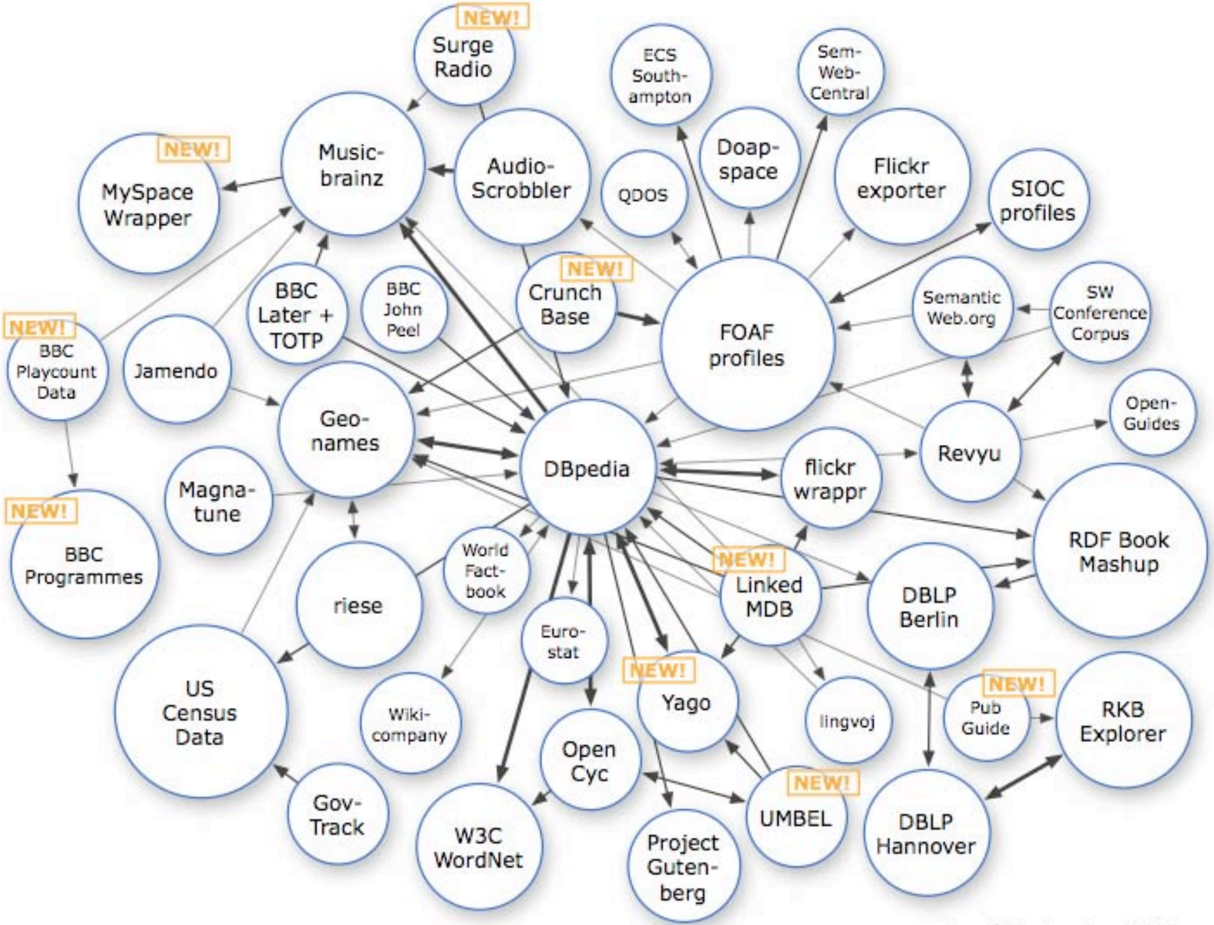
- 12 data sets
- Over 500 million RDF triples
- Around 120,000 RDF links between data sources

Web of Data (as of November 2007)



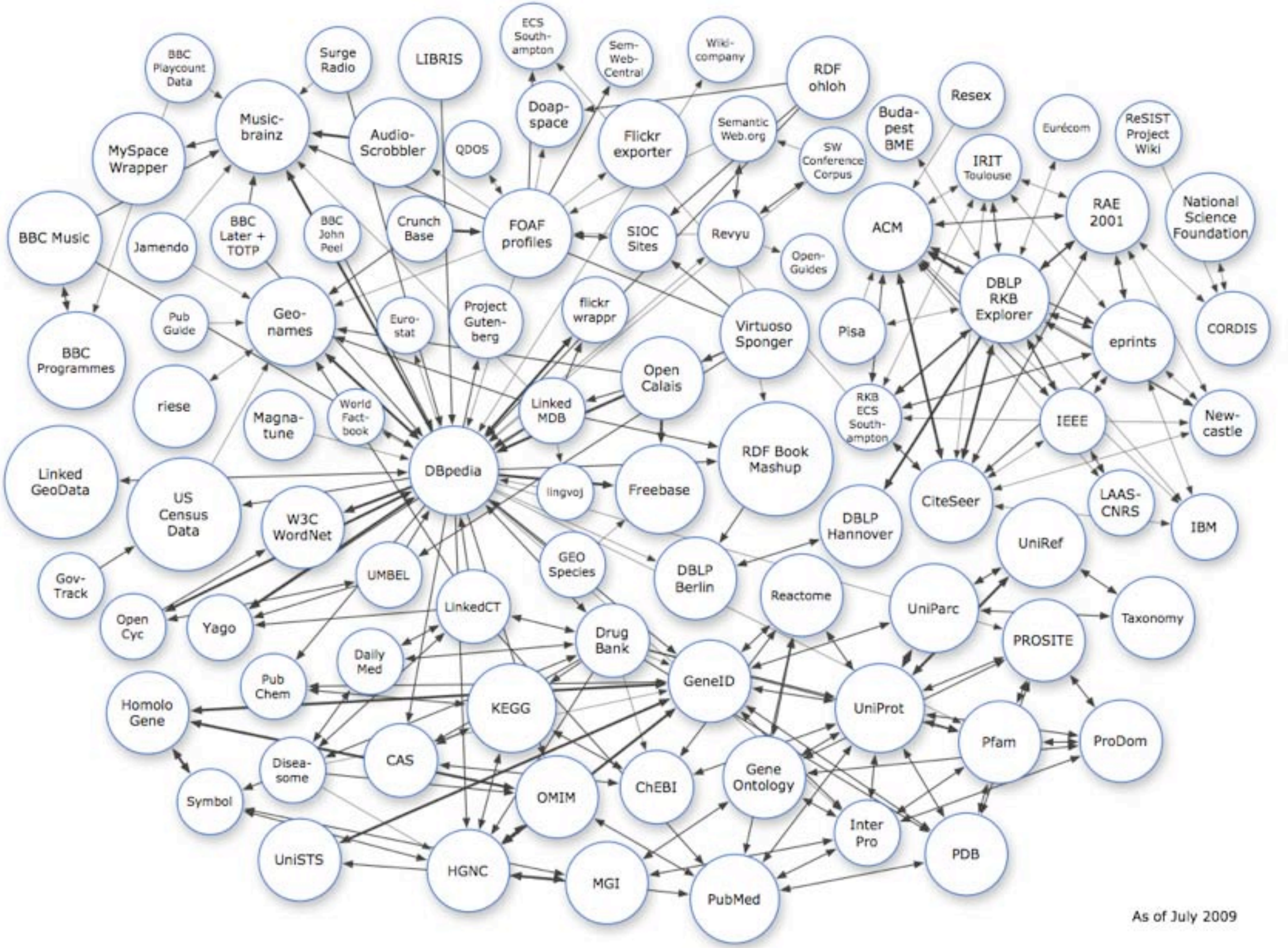
- 28 data sets

Web of Data (as of September 2008)



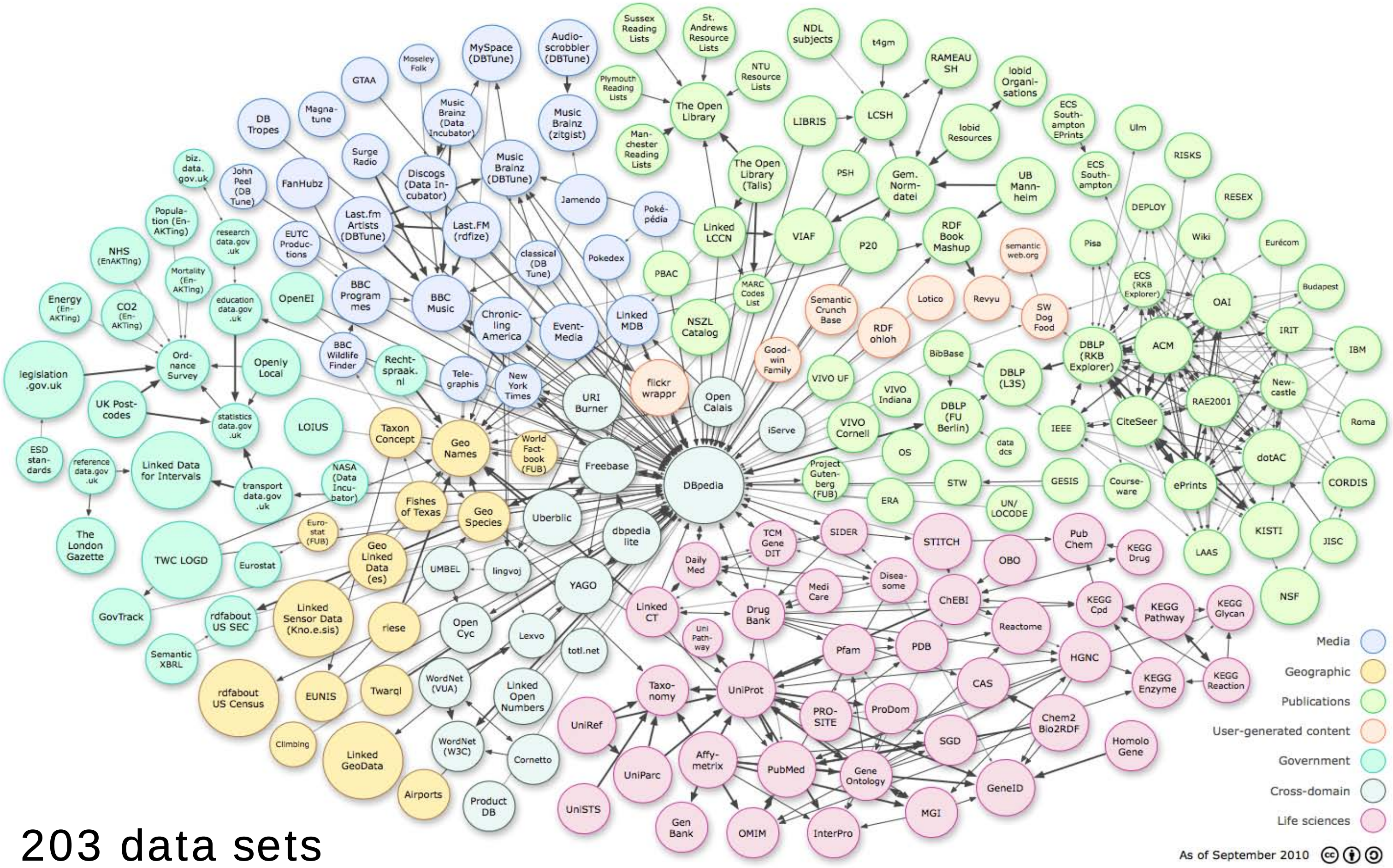
As of September 2008

Web of Data (as of July 2009)



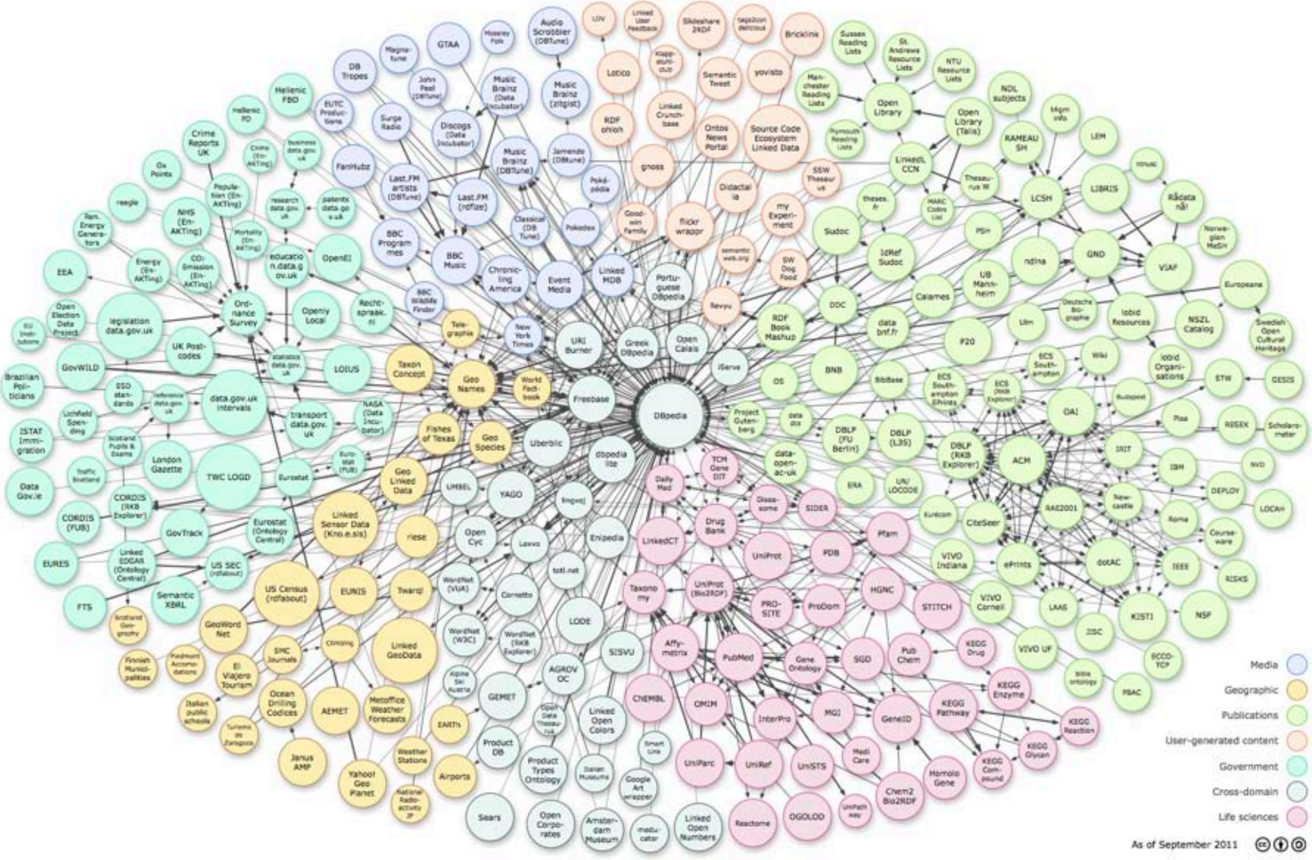
As of July 2009

Web of Data (as of September 2010)



- 203 data sets
- Over 24,7 billion RDF triples
- Over 436 million RDF links between data sources

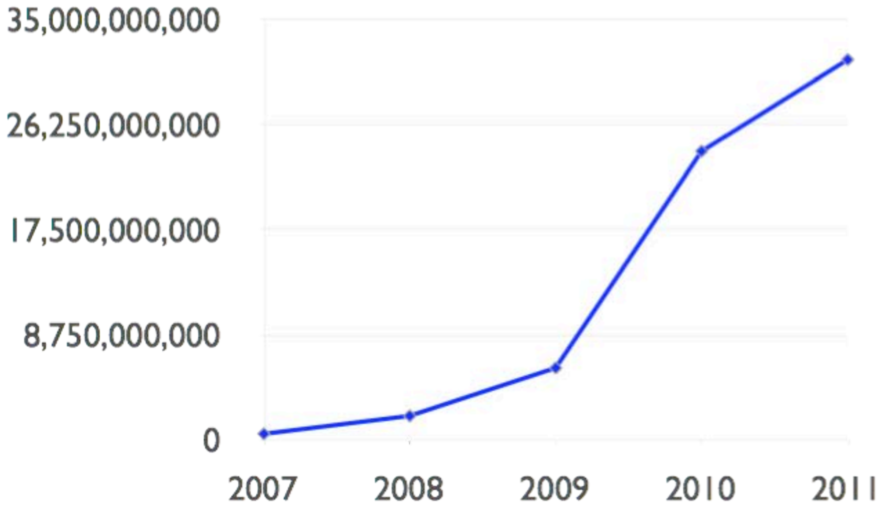
Web of Data (as of September 2011)



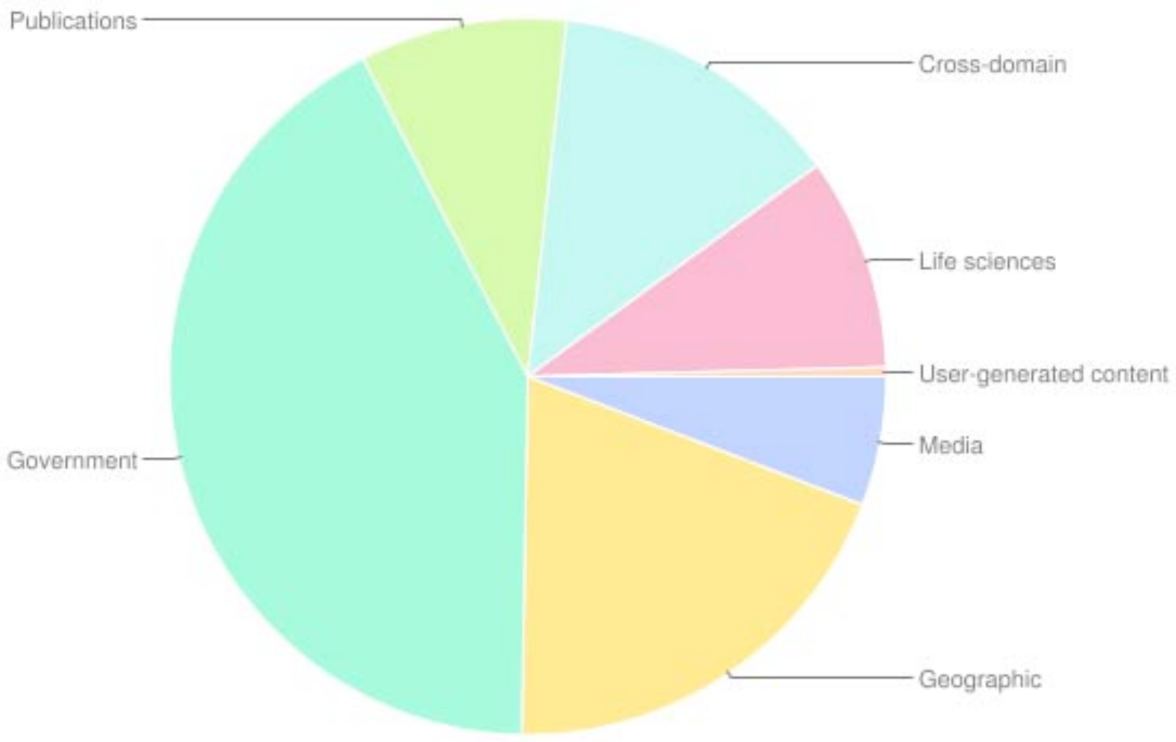
- 295 data sets
- Over 31 billion RDF triples
- Over 504 million RDF links between data sources

The Growth in Numbers

Year	Data Sets	Triples	Growth
2007	12	500,000,000	
2008	45	2,000,000,000	300%
2009	95	6,726,000,000	236%
2010	203	26,930,509,703	300%
2011	295	31,634,213,770	33%
2013	~ 900	?	?



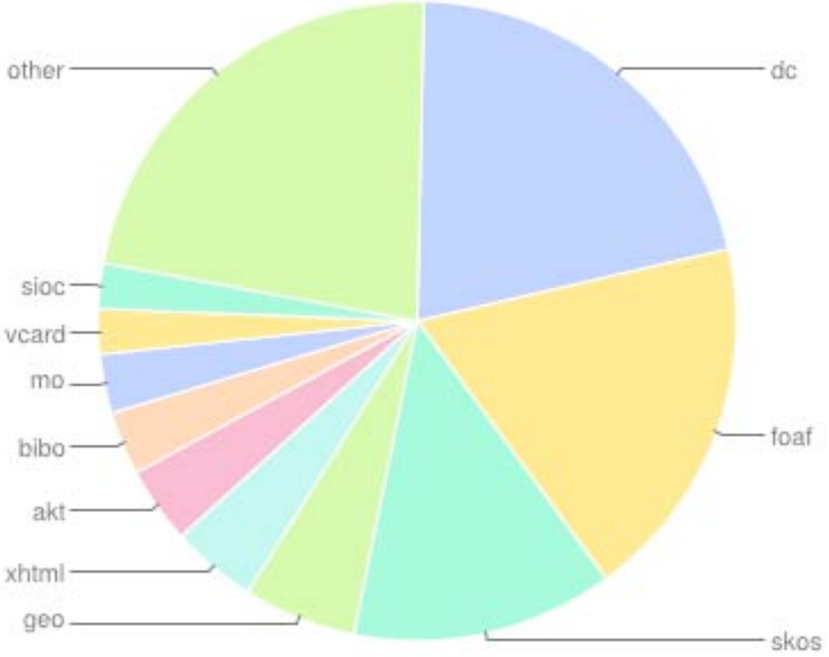
Topics on the Web of Data



- LOD Cloud Data Catalog on the Data Hub
  - □ http://datahub.io/group/lodcloud
- More statistics
  - □ http://lod-cloud.net/state/

Heterogeneity on the Web of Data

- The Web of Data is heterogeneous
  - □ Many different vocabularies are in use (337 as of April 2013)
  - □ Different data formats
  - □ Many different ways to represent the same information



Distribution of the most widely used vocabularies

Vocabularies on the Web of Data

- Common Vocabularies
  - Friend-of-a-Friend for describing people and their social network
  - SIOC for describing forums and blogs
  - SKOS for representing topic taxonomies
  - Organization Ontology for describing the structure of organizations
  - GoodRelations provides terms for describing products and business entities
  - Music Ontology for describing artists, albums, and performances
  - Review Vocabulary provides terms for representing reviews

Vocabularies on the Web of Data

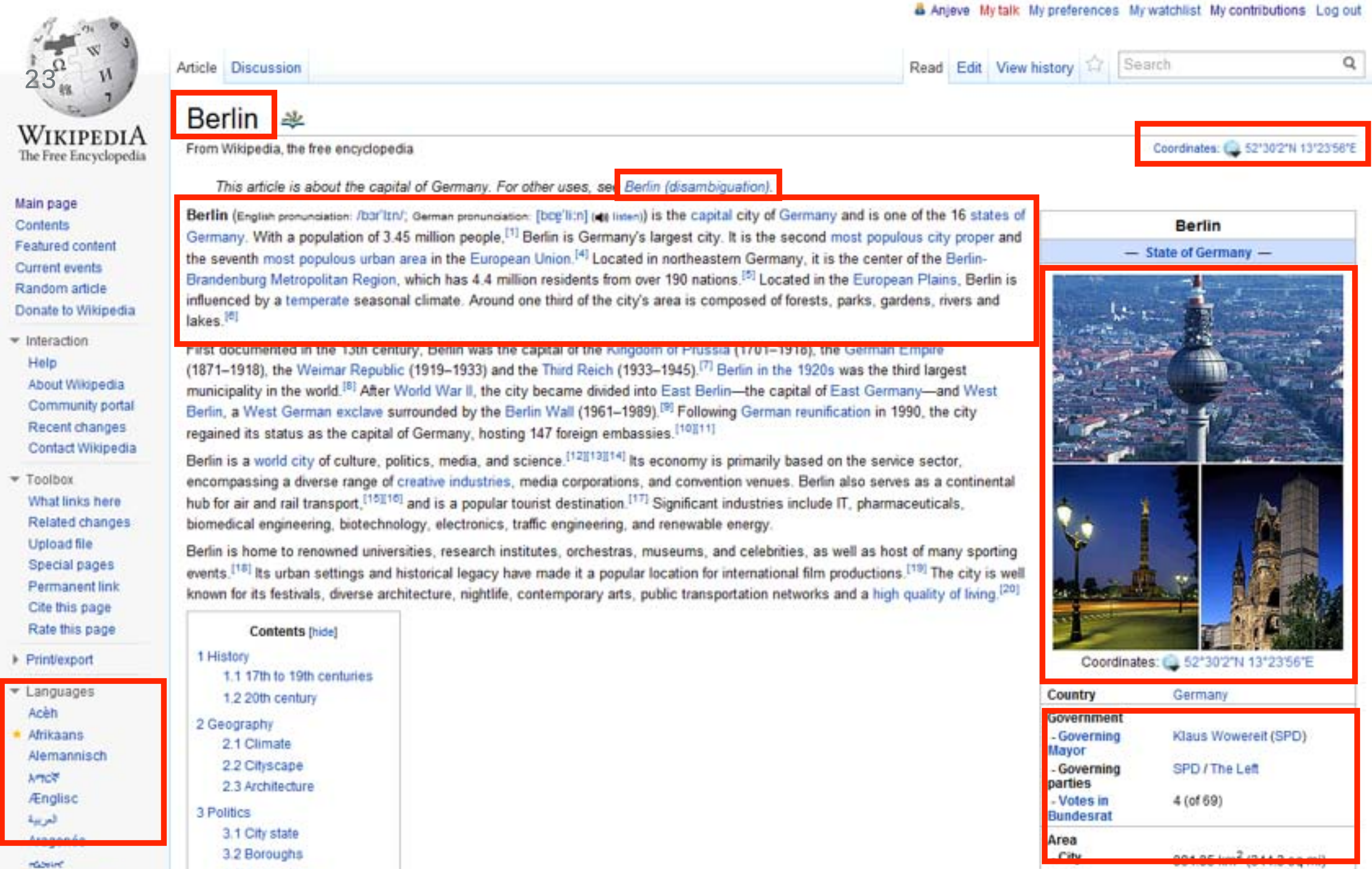
- Common sources of identifiers (URIs) for real world objects
  - LinkedGeoData and Geonames: locations
  - GeneID and UniProt: life science identifiers
  - DBpedia: wide range of things

DBpedia - The Hub on the Web of Data

- DBpedia is a joint project with the following goals
  - extracting structured information from Wikipedia
  - publish this information under an open license on the Web
  - setting links to other data sources

- Partners
  - Universität Mannheim (Germany)
  - Universität Leipzig (Germany)
  - OpenLink Software (UK)

Extracting structured data from Wikipedia



Anjeve My talk My preferences My watchlist My contributions Log out

WIKIPEDIA The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox
What links here
Related changes
Upload file
Special pages
Permanent link
Cite this page
Rate this page

Print/export

Languages
Acèh
Afrikaans
Alemannisch
አማርኛ
Ænglisc
العربية
Aragonés

Article Discussion Read Edit View history Search

Berlin

From Wikipedia, the free encyclopedia

Coordinates: 52°30′2″N 13°23′56″E

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

Berlin (English pronunciation: /bɜrˈlɪn/; German pronunciation: [bɛʁˈliːn] listen) is the capital city of Germany and is one of the 16 states of Germany. With a population of 3.45 million people,[1] Berlin is Germany's largest city. It is the second most populous city proper and the seventh most populous urban area in the European Union.[4] Located in northeastern Germany, it is the center of the Berlin-Brandenburg Metropolitan Region, which has 4.4 million residents from over 190 nations.[5] Located in the European Plains, Berlin is influenced by a temperate seasonal climate. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes.[6]

First documented in the 13th century, Berlin was 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).[7] Berlin in the 1920s was the third largest municipality in the world.[8] After World War II, the city became divided into East Berlin—the capital of East Germany—and West Berlin, a West German exclave surrounded by the Berlin Wall (1961–1989).[9] Following German reunification in 1990, the city regained its status as the capital of Germany, hosting 147 foreign embassies.[10][11]

Berlin is a world city of culture, politics, media, and science.[12][13][14] Its economy is primarily based on the service sector, encompassing a diverse range of creative industries, media corporations, and convention venues. Berlin also serves as a continental hub for air and rail transport,[15][16] and is a popular tourist destination.[17] Significant industries include IT, pharmaceuticals, biomedical engineering, biotechnology, electronics, traffic engineering, and renewable energy.

Berlin is home to renowned universities, research institutes, orchestras, museums, and celebrities, as well as host of many sporting events.[18] Its urban settings and historical legacy have made it a popular location for international film productions.[19] The city is well known for its festivals, diverse architecture, nightlife, contemporary arts, public transportation networks and a high quality of living.[20]

Contents [hide]

Berlin

— State of Germany —

Coordinates: 52°30′2″N 13°23′56″E

Country	Germany
Government	
- Governing Mayor	Klaus Wowereit (SPD)
- Governing parties	SPD / The Left
- Votes in Bundesrat	4 (of 69)
Area	
- City	891.85 km² (344.3 sq mi)

Extracting structured data from Wikipedia

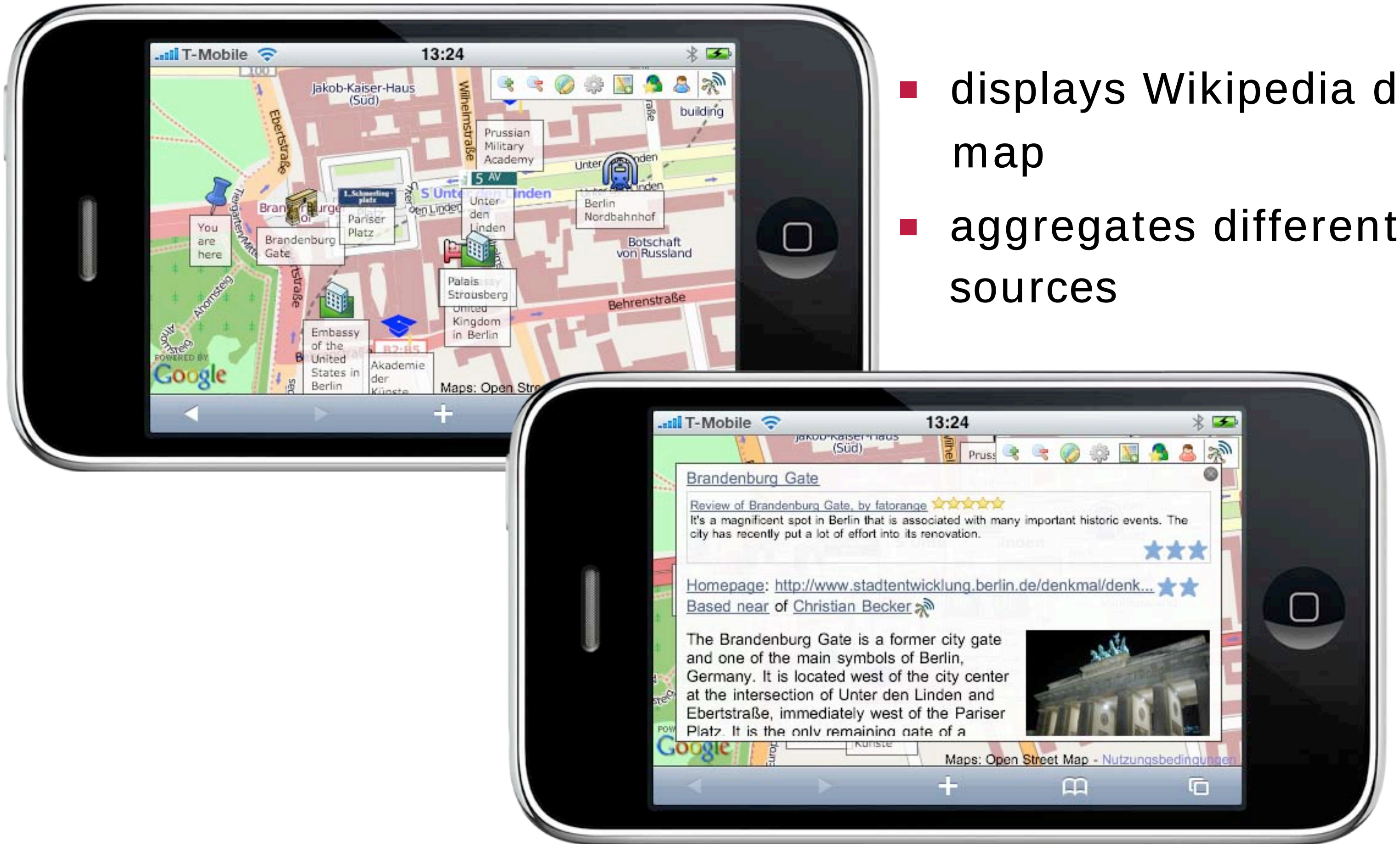
```
dbpedia:Berlin  rdf:type  dbpedia-owl:City ,
dbpedia-owl:PopulatedPlace ,
dbpedia-owl:Place ;
rdfs:label  "Berlin"@en , "Berlino"@it ;
dbpedia-owl:population  3499879 ;
wgs84:lat  52.500557 ;
wgs84:long  13.398889 .

dbpedia:SoundCloud  dbpedia-owl:location  dbpedia:Berlin .
```

- access to DBpedia data:
    - RDF dumps
    - Linked Data interface
    - SPARQL endpoint

DBpedia Use Cases

1. Improvement of Wikipedia search
2. Data source for applications and mashups
3. Text analysis and annotation
4. Hub for the growing Web of Data

DBpedia Mobile



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

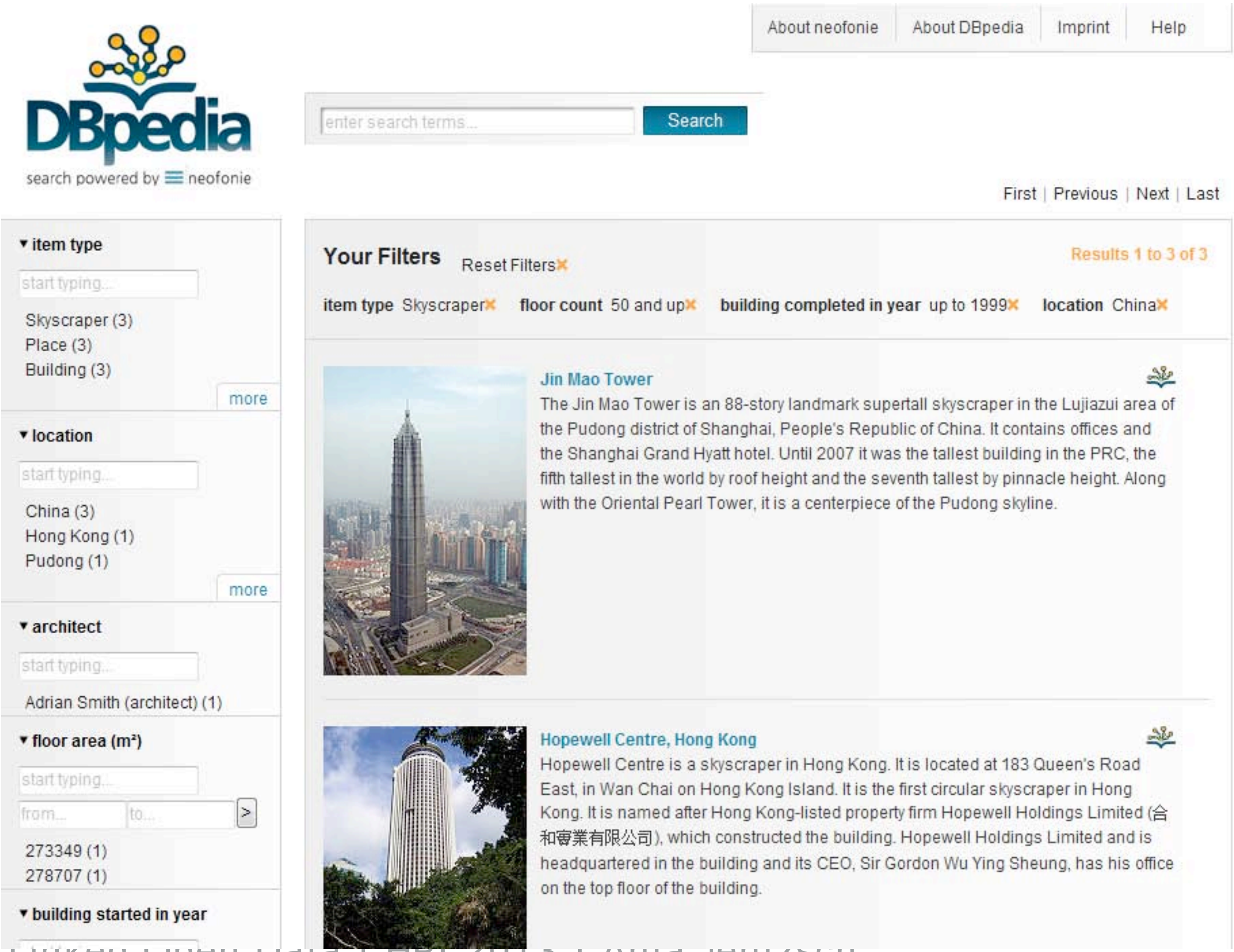
The Brandenburg Gate is a former city gate and one of the main symbols of Berlin, Germany. It is located west of the city center at the intersection of Unter den Linden and Ebertstraße, immediately west of the Pariser Platz. It is the only remaining gate of a

Maps: Open Street Map - Nutzungsbedingungen

- displays Wikipedia data on map
- aggregates different data sources

Faceted Wikipedia Search

- Faceted browsing and free text search



▾ item type

start typing...

Skyscraper (3)
Place (3)
Building (3)

more

▾ location

start typing...

China (3)
Hong Kong (1)
Pudong (1)

more

▾ architect

start typing...

Adrian Smith (architect) (1)

▾ floor area (m²)

start typing...

from... to... >

273349 (1)
278707 (1)

▾ building started in year

Your Filters Reset Filters✖ Results 1 to 3 of 3

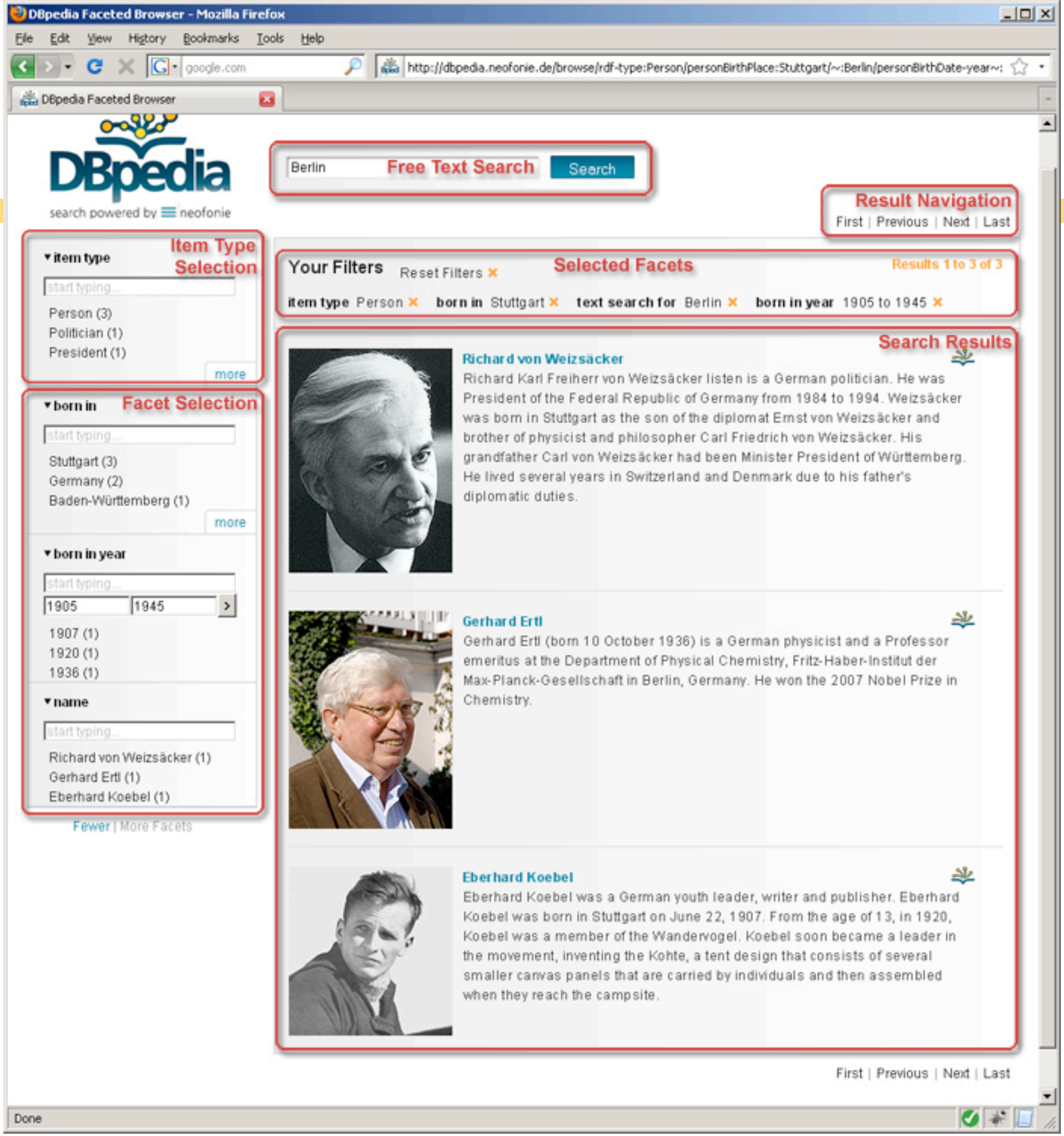
item type Skyscraper✖ **floor count** 50 and up✖ **building completed in year** up to 1999✖ **location** China✖

Jin Mao Tower

The Jin Mao Tower is an 88-story landmark supertall skyscraper in the Lujiazui area of the Pudong district of Shanghai, People's Republic of China. It contains offices and the Shanghai Grand Hyatt hotel. Until 2007 it was the tallest building in the PRC, the fifth tallest in the world by roof height and the seventh tallest by pinnacle height. Along with the Oriental Pearl Tower, it is a centerpiece of the Pudong skyline.

Hopewell Centre, Hong Kong

Hopewell Centre is a skyscraper in Hong Kong. It is located at 183 Queen's Road East, in Wan Chai on Hong Kong Island. It is the first circular skyscraper in Hong Kong. It is named after Hong Kong-listed property firm Hopewell Holdings Limited (合和實業有限公司), which constructed the building. Hopewell Holdings Limited and is headquartered in the building and its CEO, Sir Gordon Wu Ying Sheung, has his office on the top floor of the building.


DBpedia Faceted Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
google.com
http://dbpedia.neofonie.de/browse/rdf-type:Person/personBirthPlace:Stuttgart/~:Berlin/personBirthDate-year~:
DBpedia Faceted Browser
DBpedia
search powered by neofonie
Free Text Search
Search
Result Navigation
First | Previous | Next | Last
Item Type Selection
Facet Selection
Selected Facets
Search Results
more
Fewer | More Facets
Done

Berlin

Your Filters Reset Filters ✕ Results 1 to 3 of 3

item type Person ✕ born in Stuttgart ✕ text search for Berlin ✕ born in year 1905 to 1945 ✕

▾ item type

start typing...

Person (3)
Politician (1)
President (1)

▾ born in

start typing...

Stuttgart (3)
Germany (2)
Baden-Württemberg (1)

▾ born in year

start typing...

1905 1945 >

1907 (1)
1920 (1)
1936 (1)

▾ name

start typing...

Richard von Weizsäcker (1)
Gerhard Ertl (1)
Eberhard Koebel (1)

Richard von Weizsäcker

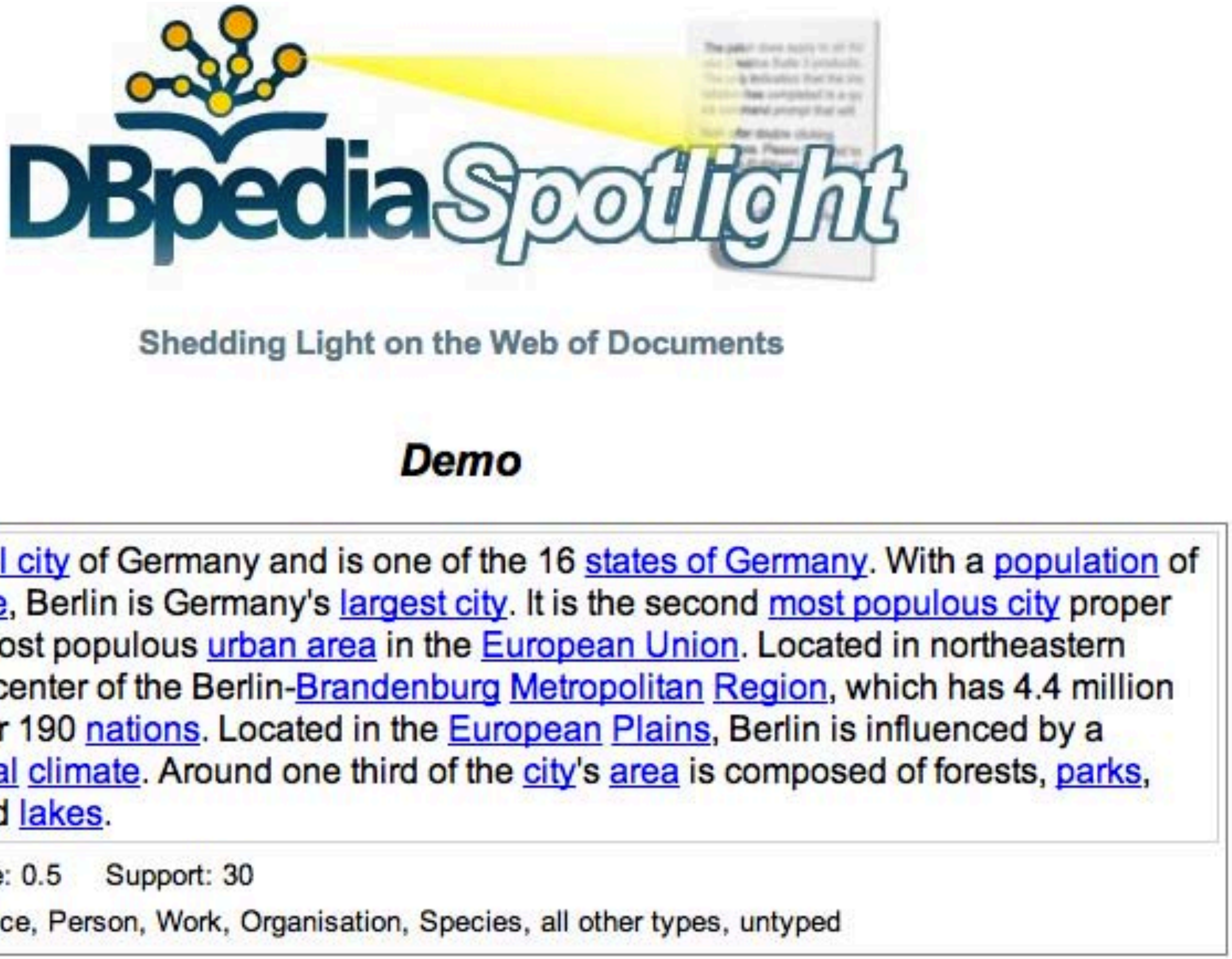
Richard Karl Freiherr von Weizsäcker listen is a German politician. He was President of the Federal Republic of Germany from 1984 to 1994. Weizsäcker was born in Stuttgart as the son of the diplomat Ernst von Weizsäcker and brother of physicist and philosopher Carl Friedrich von Weizsäcker. His grandfather Carl von Weizsäcker had been Minister President of Württemberg. He lived several years in Switzerland and Denmark due to his father's diplomatic duties.

Gerhard Ertl

Gerhard Ertl (born 10 October 1936) is a German physicist and a Professor emeritus at the Department of Physical Chemistry, Fritz-Haber-Institut der Max-Planck-Gesellschaft in Berlin, Germany. He won the 2007 Nobel Prize in Chemistry.

Eberhard Koebel

Eberhard Koebel was a German youth leader, writer and publisher. Eberhard Koebel was born in Stuttgart on June 22, 1907. From the age of 13, in 1920, Koebel was a member of the Wandervogel. Koebel soon became a leader in the movement, inventing the Kohte, a tent design that consists of several smaller canvas panels that are carried by individuals and then assembled when they reach the campsite.



Shedding Light on the Web of Documents

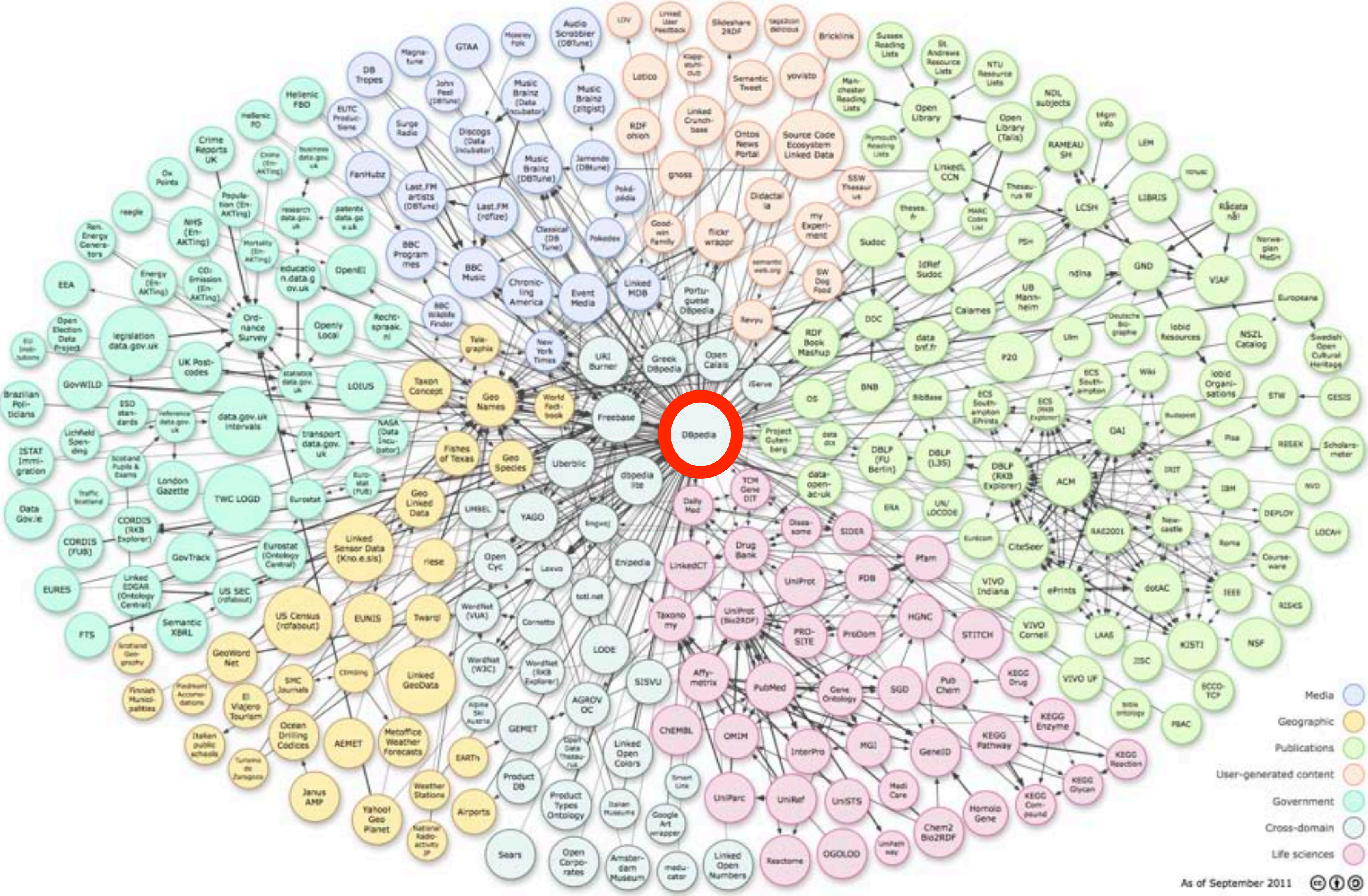
Demo

Berlin is the capital city of Germany and is one of the 16 states of Germany. With a population of 3.45 million people, Berlin is Germany's largest city. It is the second most populous city proper and the seventh most populous urban area in the European Union. Located in northeastern Germany, it is the center of the Berlin-Brandenburg Metropolitan Region, which has 4.4 million residents from over 190 nations. Located in the European Plains, Berlin is influenced by a temperate seasonal climate. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes.

Confidence: 0.5 Support: 30

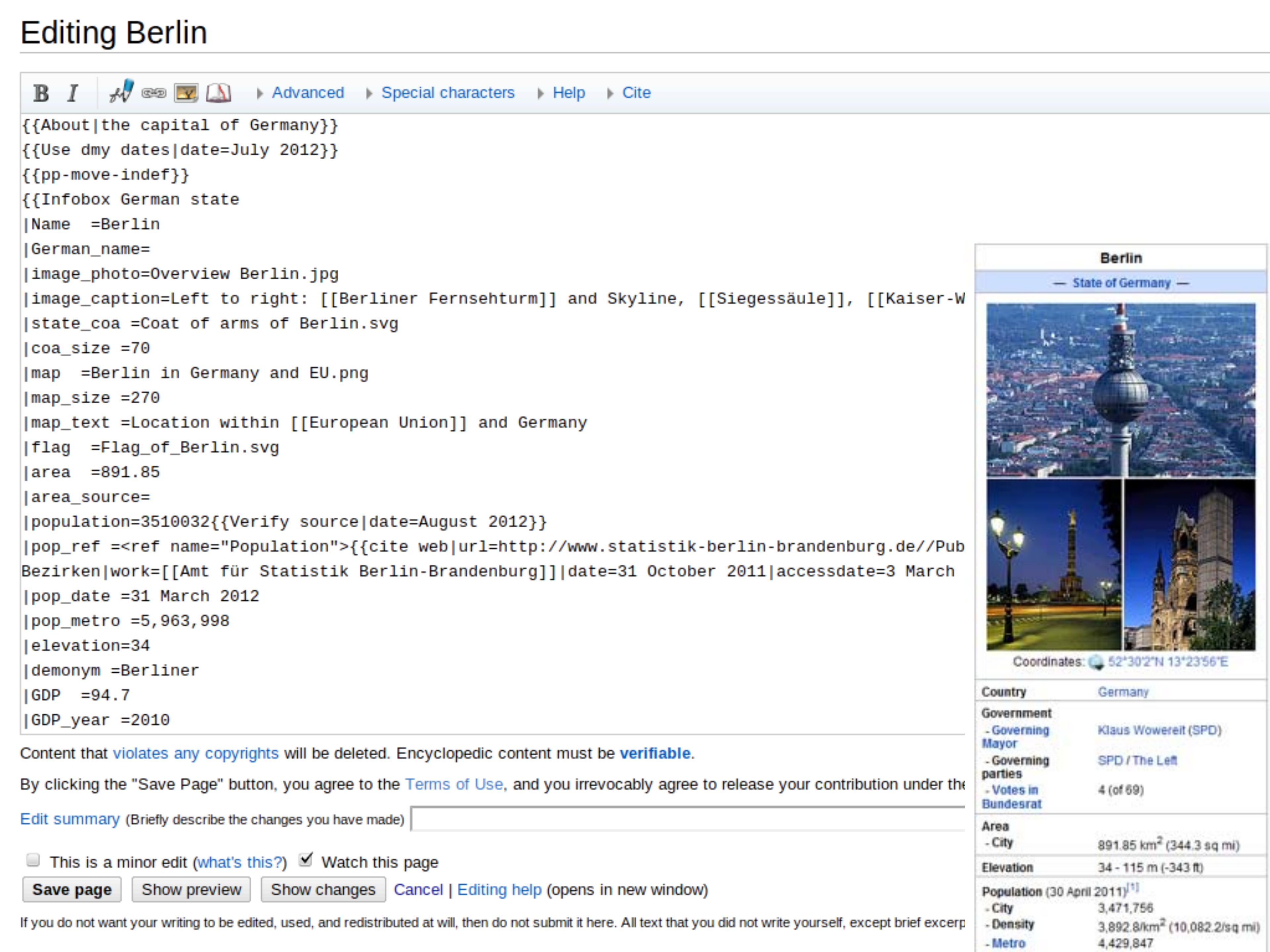
Types: Place, Person, Work, Organisation, Species, all other types, untyped

http://spotlight.dbpedia.org


DBpedia
Freebase
Geo Names
YAGO
UMBEL
Uberblic
dbpedia lite
Open Cyc
LinkedMDB
MusicBrainz (zitgist)
MusicBrainz (DBTune)
Music Brainz (Data Incubator)
BBC Music
BBC Program mes
BBC Wildlife Finder
Last.FM (rdfize)
Last.FM artists (DBTune)
Discogs (Data Incubator)
Jamendo (DBtune)
Audio Scrobbler (DBTune)
GTAA
DB Tropes
Magnatune
John Peel (DBTune)
Moseley Folk
Surge Radio
FanHubz
Classical (DB Tune)
Pokedex
Event Media
Chronicling America
New York Times
Telegraphis
Taxon Concept
World Factbook
Fishes of Texas
Geo Species
Geo Linked Data
Linked Sensor Data (Kno.e.sis)
riese
US Census (rdfabout)
EUNIS
Twarql
GeoWord Net
Climbing
Linked GeoData
SMC Journals
El Viajero Tourism
Ocean Drilling Codices
AEMET
Metoffice Weather Forecasts
EARTH
Airports
Weather Stations
Janus AMP
Yahoo! Geo Planet
National Radioactivity JP
Alpine Ski Austria
Scotland Geography
Finnish Municipalities
Piedmont Accomodations
Italian public schools
Turismo de Zaragoza
UK Postcodes
legislation data.gov.uk
Ordnance Survey
OpenEI
Openly Local
educatio n.data.g ov.uk
Crime Reports UK
Hellenic FBD
Hellenic PD
EUTC Productions
reegle
Ox Points
NHS (En-AKTing)
Population (En-AKTing)
Crime (En-AKTing)
business data.gov.uk
research data.gov.uk
patents data.go v.uk
Mortality (En-AKTing)
Energy (En-AKTing)
CO2 Emission (En-AKTing)
Ren. Energy Generators
EEA
Open Election Data Project
EU Institutions
GovWILD
Brazilian Politicians
ESD standards
reference data.gov.uk
data.gov.uk intervals
statistics data.gov.uk
LOIUS
transport data.gov.uk
NASA (Data Incubator)
Recht-spraak.nl
Lichfield Spending
ISTAT Immigration
Scotland Pupils & Exams
London Gazette
TWC LOGD
Eurostat
Euro-stat (FUB)
Traffic Scotland
Data Gov.ie
CORDIS (RKB Explorer)
CORDIS (FUB)
GovTrack
Eurostat (Ontology Central)
EURES
Linked EDGAR (Ontology Central)
US SEC (rdfabout)
Semantic XBRL
FTS
UK Burner
Greek DBpedia
Open Calais
iServe
Portuguese DBpedia
Revyu
LOV
Linked User Feedback
Slideshare 2RDF
tags2con delicious
Bricklink
Lotico
Klappstuhlclub
Semantic Tweet
yovisto
RDF ohloh
Linked Crunchbase
Ontos News Portal
Source Code Ecosystem Linked Data
gnoss
Pokepedia
Didactalia
Goodwin Family
flickr wrappr
my Experiment
semantic web.org
SW Dog Food
SSW Thesaur us
OS
Project Gutenberg
data dcs
DBLP (FU Berlin)
DBLP (L3S)
DBLP (RKB Explorer)
data-open-ac-uk
ERA
UN/ LOCODE
TCM Gene DIT
Daily Med
Diseasome
SIDER
Drug Bank
LinkedCT
Pfam
UniProt
PDB
HGNC
STITCH
UniProt (Bio2RDF)
Taxonomy
PRO-SITE
ProDom
Affymetrix
PubMed
Gene Ontology
SGD
Pub Chem
KEGG Drug
KEGG Enzyme
KEGG Pathway
KEGG Reaction
KEGG Glycan
KEGG Compound
ChEMBL
OMIM
InterPro
MGI
GeneID
UniParc
UniRef
UniSTS
Medi Care
Homolo Gene
Chem2 Bio2RDF
Reactome
OGOLOD
UniPath way
Linked Open Numbers
Linked Open Colors
Product DB
Product Types Ontology
Open Corporates
Sears
Amsterdam Museum
Italian Museums
Google Art wrapper
meducator
Smart Link
Open Data Thesaurus
GEMET
AGROV OC
LODE
SISVU
totl.net
Enipedia
Lexvo
Cornetto
lingvoj
WordNet (VUA)
WordNet (W3C)
WordNet (RKB Explorer)
Sussex Reading Lists
St. Andrews Resource Lists
NTU Resource Lists
Manchester Reading Lists
Plymouth Reading Lists
Open Library
Open Library (Talis)
NDL subjects
Mgm info
LEM
RAMEAU SH
Linked CCN
Thesaurus W
theses.fr
MARC Codes List
LCSH
LIBRIS
ndlna
Rådata nå!
Norwegian MeSH
Sudoc
IdRef Sudoc
PSH
UB Mannheim
GND
VIAF
Europeana
DDC
Calames
Deutsche Biographie
lobid Resources
NSZL Catalog
Swedish Open Cultural Heritage
RDF Book Mashup
data bnf.fr
P20
Ulm
ECS Southampton
Wiki
lobid Organisations
STW
GESIS
BNB
BibBase
ECS Southampton EPrints
ECS (RKB Explorer)
Budapest
OAI
Pisa
RESEX
Scholarometer
IRIT
IBM
NVD
ACM
DEPLOY
LOCAH
RAE2001
Newcastle
Roma
Courseware
Eurécom
CiteSeer
Pfam
VIVO Indiana
ePrints
dotAC
IEEE
RISKS
VIVO Cornell
LAAS
KISTI
NSF
JISC
VIVO UF
ECCO-TCP
bible ontology
PBAC
Media
Geographic
Publications
User-generated content
Government
Cross-domain
Life sciences
As of September 2011

The DBpedia Data Set

- Information on more than 3.77 million “things”
  - □ 764,000 persons
  - □ 192,000 organisations
  - □ 573,000 places
  - □ 112,000 music albums
  - □ 72,000 movies
  - □ 202,000 species
- overall more than 1 billion RDF triples
  - □ title and abstract in 111 different languages
  - □ 8,000,000 links to images
  - □ 24,400,000 links to external web pages
  - □ 27,200,000 links to other Linked Data sets



Editing Berlin

```
{{About|the capital of Germany}}
{{Use dmy dates|date=July 2012}}
{{pp-move-indef}}
{{Infobox German state
|Name  =Berlin
|German_name=
|image_photo=Overview Berlin.jpg
|image_caption=Left to right: [[Berliner Fernsehturm]] and Skyline, [[Siegessäule]], [[Kaiser-W
|state_coa =Coat of arms of Berlin.svg
|coa_size =70
|map  =Berlin in Germany and EU.png
|map_size =270
|map_text =Location within [[European Union]] and Germany
|flag  =Flag_of_Berlin.svg
|area  =891.85
|area_source=
|population=3510032{{Verify source|date=August 2012}}
|pop_ref =<ref name="Population">{{cite web|url=http://www.statistik-berlin-brandenburg.de//Pub
Bezirken|work=[[Amt für Statistik Berlin-Brandenburg]]|date=31 October 2011|accessdate=3 March
|pop_date =31 March 2012
|pop_metro =5,963,998
|elevation=34
|demonym =Berliner
|GDP  =94.7
|GDP_year =2010
```

Content that violates any copyrights will be deleted. Encyclopedic content must be **verifiable**.

By clicking the "Save Page" button, you agree to the Terms of Use, and you irrevocably agree to release your contribution under the

Edit summary (Briefly describe the changes you have made)

☐ This is a minor edit (what's this?) ☑ Watch this page
Save page
Show preview
Show changes
Cancel | Editing help (opens in new window)

If you do not want your writing to be edited, used, and redistributed at will, then do not submit it here. All text that you did not write yourself, except brief excerp

Berlin

— State of Germany —

Coordinates: 52°30′2″N 13°23′56″E

Country	Germany
Government	
- Governing Mayor	Klaus Wowereit (SPD)
- Governing parties	SPD / The Left
- Votes in Bundesrat	4 (of 69)
Area	
- City	891.85 km^2 (344.3 sq mi)
Elevation	34 - 115 m (-343 ft)
Population (30 April 2011)[1]	
- City	3,471,756
- Density	3,892.8/km^2 (10,082.2/sq mi)
- Metro	4,429,847

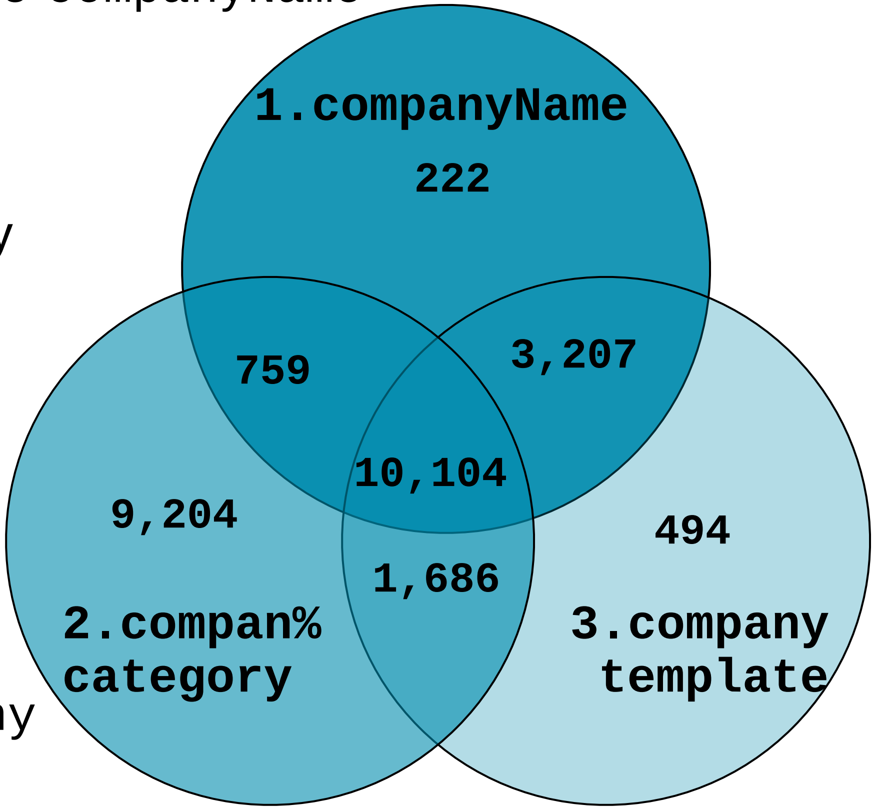
Linked Data Heterogeneity

- Companies in DBpedia
- Def. 1: Subject having a predicate companyName

  → 14,292
- Def. 2: Subject having a category that starts with 'compan'

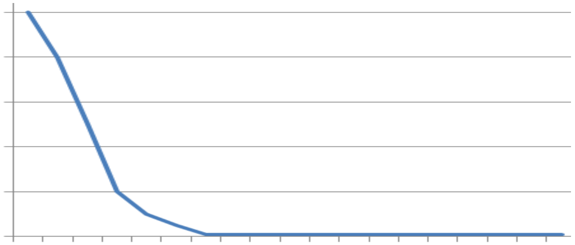
  → 21,753
- Def. 3: Subject having a wikiPageUsesTemplate with value Template:infobox_company

  → 15,491



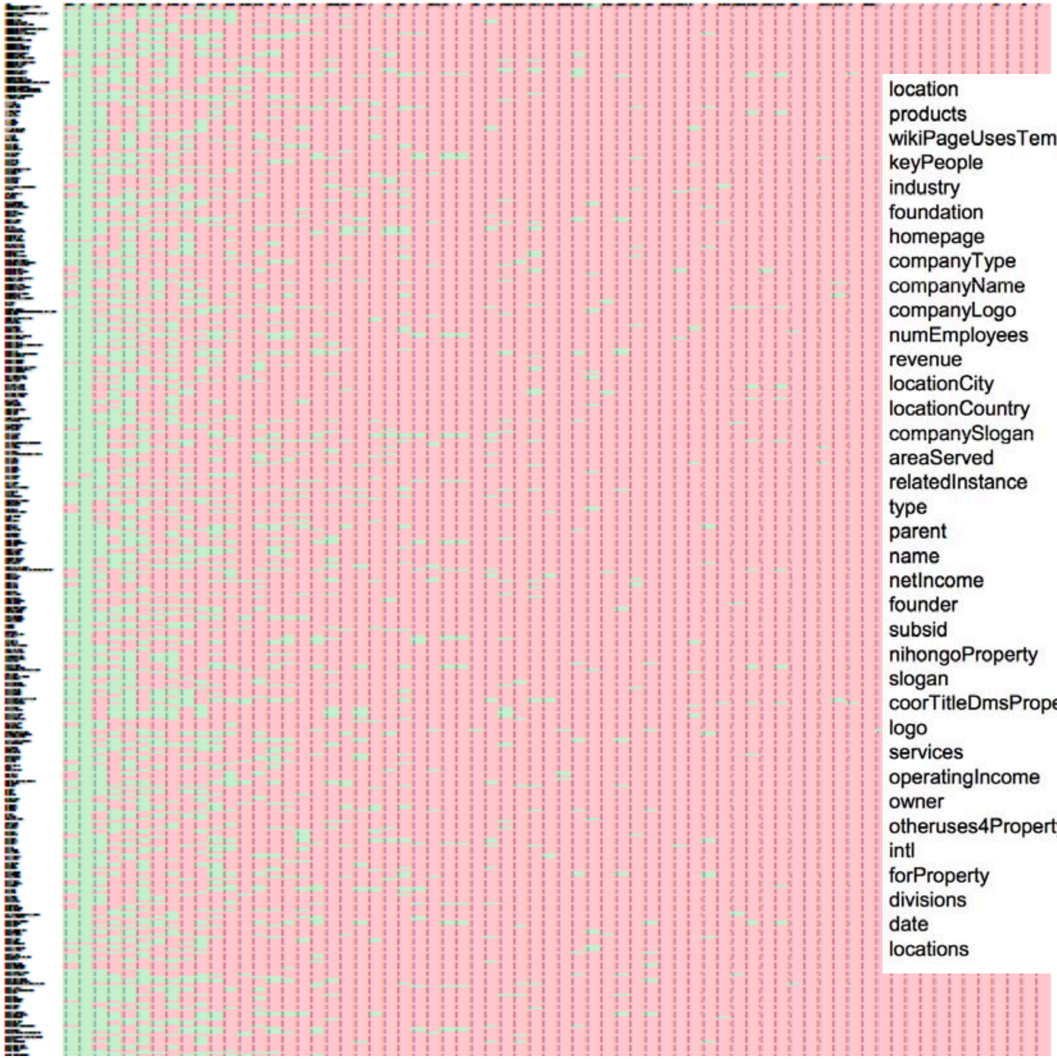
Linked Data Heterogeneity

- DBpedia: `?c wikiPageUsesTemplate Template:infobox_company`
- 1,083 different attributes
- 499 appear only once



- 39 distinct ones contain name as substring
  `companyName, commonName, publicName, …`
- 273 companies without any name attribute

DBpedia Company Attribute Distribution



location	20617	companyName	13355
products	18176	name	2036
wikiPageUsesTemplate	18048	surname	25
keyPeople	17836	railroadName	8
industry	16822	companyNickname	4
foundation	15826	pastNames	4
homepage	14476	absNameProperty	3
companyType	13433	dnvNameProperty	3
companyName	13355	labelName	3
companyLogo	9006	logoFilename	3
numEmployees	6207	dvdEuroCompanyName	2
revenue	5030	filename	2
locationCity	4098	longName	2
locationCountry	3212	websitename	2
companySlogan	2815	alternativeNames	1
areaServed	2557	birthname	1
relatedInstance	2284	brandName	1
type	2152	bTcgvuvCompanyName	1
parent	2054	companyNameLocal	1
name	2036	companyNamesBigBum	1
netIncome	1663	europeanTradeAssociationCompanyName	1
founder	1597	familyCorporationCompanyName	1
subsid	1232	formerNames	1
nihongoProperty	1141	fukCompanyName	1
slogan	1087	golfFacilityName	1
coorTitleDmsProperty	960	hangulName	1
logo	925	iceCreamCompanyName	1
services	904	nativeName	1
operatingIncome	896	nickname	1
owner	680	officialName	1
otheruses4Property	510	oldName	1
intl	503	organisationName	1
forProperty	467	publicCompanyName	1
divisions	429	renamed	1
date	422	shortName	1
locations	419	wineryName	1

DBpedia Mappings

- since March 2010 collaborative editing of
  - DBpedia ontology
  - mappings from Wikipedia infoboxes and tables to DBpedia ontology
- curated in a public wiki with instant validation methods
  - http://mappings.dbpedia.org
- multi-langual mappings to the DBpedia ontology:
  - ar, bg, bn, ca, cs, de, el, en, es, et, eu, fr, ga, hi, hr, hu, it, ja, ko, nl, pl, pt, ru, sl, tr
- allows for a significant increase of the extracted data’s quality
  - each domain has its experts

DBpedia Ontology

- 359 classes
  - 2,347 mappings from Wikipedia infoboxes to ontology classes (overall)
- 800 object properties, 859 datatype properties, 116 specialized datatype properties
  - 5,859 mappings from Wikipedia infobox properties to ontology properties (en)
- 45 owl:equivalentClass and 31 owl:equivalentProperty mappings to http://schema.org

Linked Data Schema Mess

- Example: Wikipedia/DBpedia
- Schema chaos: Many attribute synonyms
  - Hundreds of different attributes
  - `companyName` vs. `organizationName` vs. `name` vs. `company`
- Schema misuse: Many attribute homonyms
  - `foundation` attribute in DBPedia may contain
    - Person who founded the company
    - Year/Date company was founded
    - Location where the company was found

Linked Data Schema Mess

- Linked Data published by third parties
  - Personal view on data
  - Misinterpretation

- Loosely defined schema
  - Missing property definitions
  - Property types used inconsistently

Outline

- Introduction to Linked Data
  - □ Data Model
  - □ Data Variety
  - □ Example Data Set: DBpedia
- Profiling Linked Data
  - □ Challenges
  - □ Comparison: Traditional vs Linked Data Profiling
  - □ Existing Approaches

Profiling Linked Data - Motivation

- Current situation:
  - Web of Data is growing
- Advantages:
  - Wealth of information
  - Easy, public access
  - Interesting domains

Profiling Linked Data - Motivation

- Challenges:
  - Heterogeneity

◇ Loose structure	Things have different predicate stets
◇ Incomplete	Subjects do not have name predicate
◇ Poorly formatted	Predicate values have many patterns
◇ Inconsistent	Multiple representations claim opposite

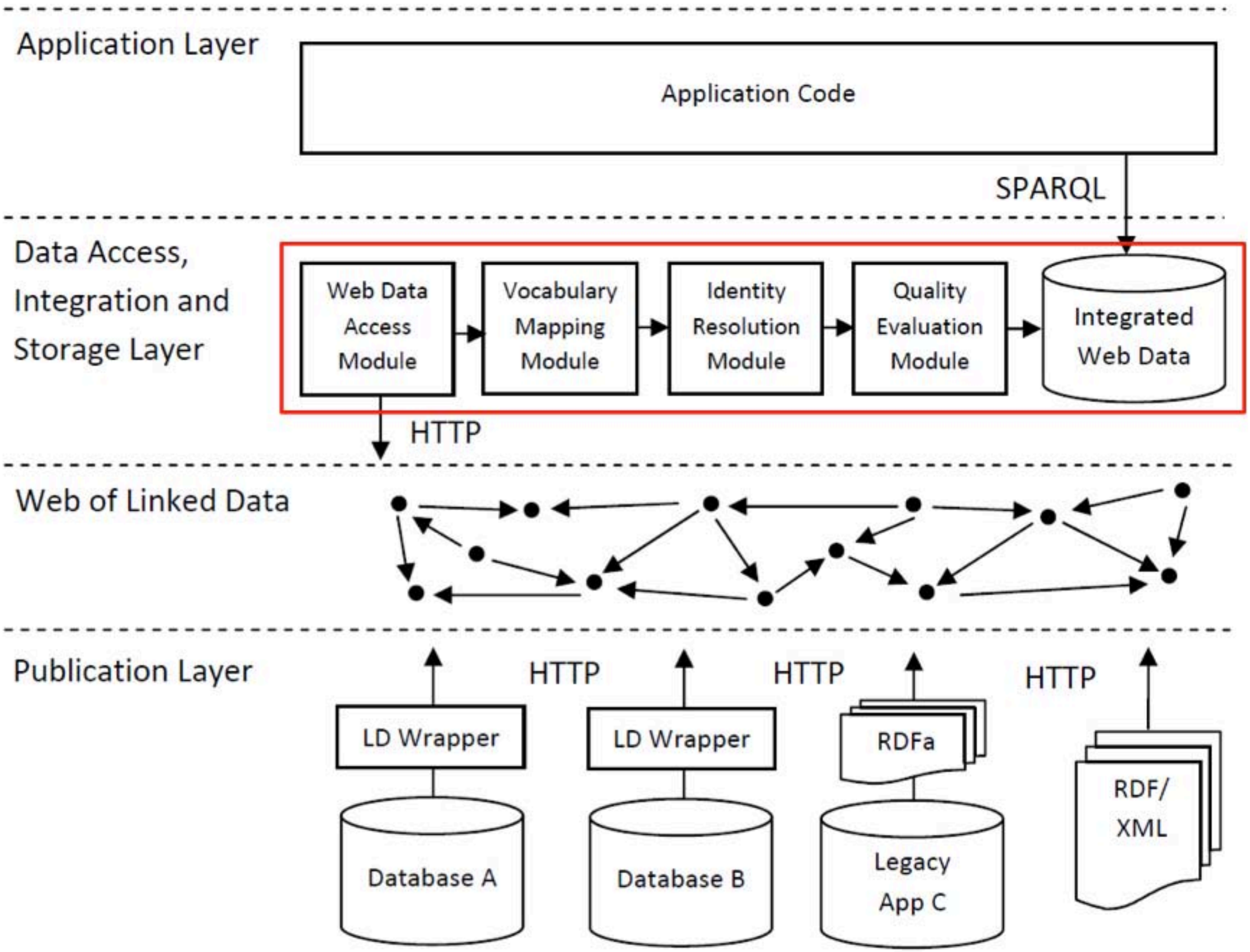
  - Volume of data

Profiling Linked Data - Use Cases

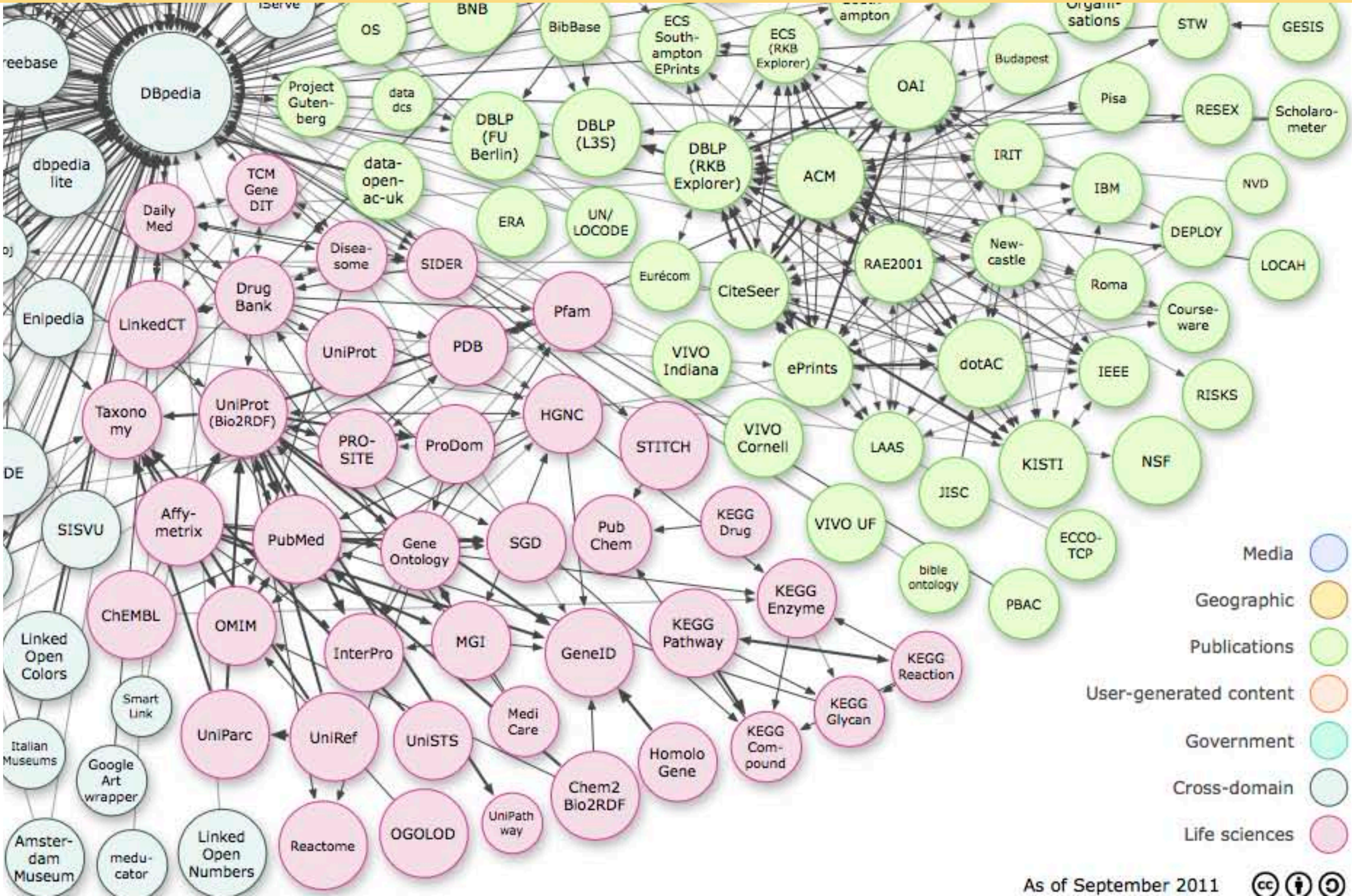
- Linked Data integration
- Linked Data publication
- Interlinking Linked Data sets

- Data profiling allows for analyzing
  - Semantic heterogeneity
  - Structural heterogeneity

Linked Data Integration Process



Example: Life Sciences



Describing Linked Data Sets

- Required knowledge for describing Linked Data sets:
  - Detailed characteristics of a data set (or parts of it)
  - Relevance of data set
  - Retrieving and processing these information for a large number of data sets is practically unfeasable
- Easy finding approach:
  - Popular data sets (e.g. DBpedia, Geonames)
  - Not always optimal:
    - If data domain is highly specialized and not covered by popular data sets in sufficient detail
    - If different parts of the data sets are covered by several external data sets (e.g. publications both on computer science (DBLP) and medicine (PubMed))

Profiling Linked Data - Motivation

- Evolving Linked Data sets require constant re-analysis
- Interlinking Linked Data sets
  - □ Link discovery problem has been addressed by several approaches (Silk, LIMES, KnoFuss)
  - □ Published data sets often interlinked with the help of researchers interested in the Linked Data initiative
- Identifying relevant sources did not acquire much attention
- Gathering linkage/integration possibilities is a time-consuming effort
- Reduce effort to perform exploratory search
- Bringing publication and interlinking process closer together

What describes a Linked Data Set?

- Topic(s)
- Statistical characteristics
  - Classes
  - Instances
  - Properties
  - Property values (and distribution)
- Language(s)
- Schema
- Data set granularity
- Relevance
- …

Where to find Information on a Data Set?

- Documents on the data set (website, papers, ...)
- Metadata files (VoID / Semantic Sitemap)
- Data registries (The Data Hub)

- Provide valuable but usually not fine-grained information on content of Linked Data sets

Traditional vs Linked Data Profiling

- State of the art Data profiling
  - Based on columns
  - Assumes well-defined semantics
  - Expects regular data

- Heterogeneity on the Web of Data
  - Diverse sources
  - ➡ Diverse structures
  - ➡ Diverse views
- RDF: nested graphs

- Nevertheless some “clean” LOD sources exist (ontologies, RDFS)
- Integration problem remains

Data Set Statistics: Instance-Based

- Number of Triples
- Number of Instances
- Average number of properties per instance

Data Set Statistics: Schema-Based

- Number of classes
- Number of instances per class
- Average number of values per property
- Percentage of top-k properties per class
- Number of different datatypes and language tags used
- Average length of strings (per property)
- Value ranges for numeric properties (per property)
- Ratio URIs/literals as objects
- Co-occuring classes
- Co-occuring properties
- Equivalent classes
- Equivalent properties

Data Set Statistics: Data Set-Based

- Number of different properties per data set and class
- Number of RDF links set between instances of the data set
- Number of RDF links pointing at instances within the dataset
- Number of RDF links pointing at instances in other data setsAverage indegree/outdegree
- Number of links likely pointing at HTML pages

Data Set Statistics: External Data-Based

- Number of classes/properties that are reused from common vocabularies
- Percentage of classes/properties that are reused from common vocabularies
- Topic (VoID, Semantic Sitemaps, The Data Hub, ...)

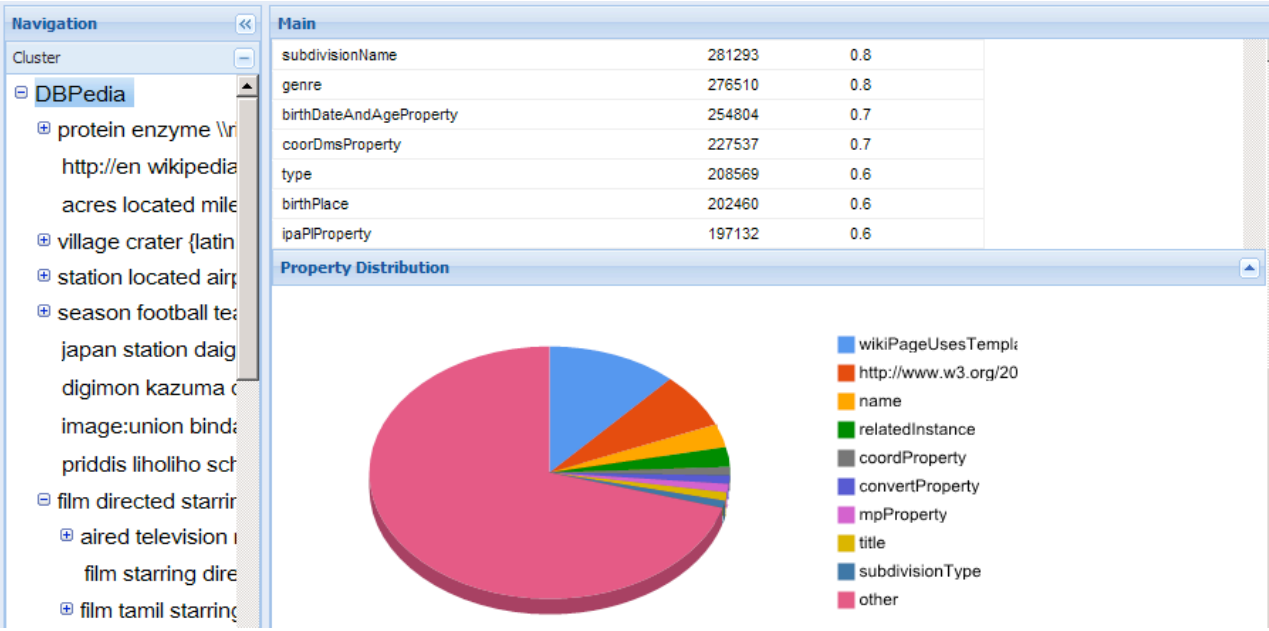
Existing Linked Data Profiling Approaches

- ProLOD
- Creating voiD descriptions
- Finding relevant link target
- Schema induction (gold-miner)

ProLOD

- Christoph Böhm, Felix Naumann et. al. @ NTII2010, ICDE2010

- Offers profiling methods to deal with loosely structured, unclean and inconsistent data on the Web of Data
- Well-known profiling techniques
- Web-based tool



Navigation

Cluster

- ⊟ DBPedia
  - ⊞ protein enzyme \\r
  - http://en wikipedia
  - acres located mile
  - ⊞ village crater {latin
  - ⊞ station located airp
  - ⊞ season football tea
  - japan station daig
  - digimon kazuma c
  - image:union binda
  - priddis liholiho sch
  - ⊟ film directed starrir
    - ⊞ aired television r
    - film starring dire
    - ⊞ film tamil starring

Main

subdivisionName	281293	0.8
genre	276510	0.8
birthDateAndAgeProperty	254804	0.7
coorDmsProperty	227537	0.7
type	208569	0.6
birthPlace	202460	0.6
ipaPlProperty	197132	0.6

Property Distribution

ProLOD

- Suite of methods ranging from:
    - Domain level (clustering, labeling)
    - Schema level (matching, disambiguation)
    - Data level (data type detection, pattern detection, value distribution)

ProLOD - Data Partitioning

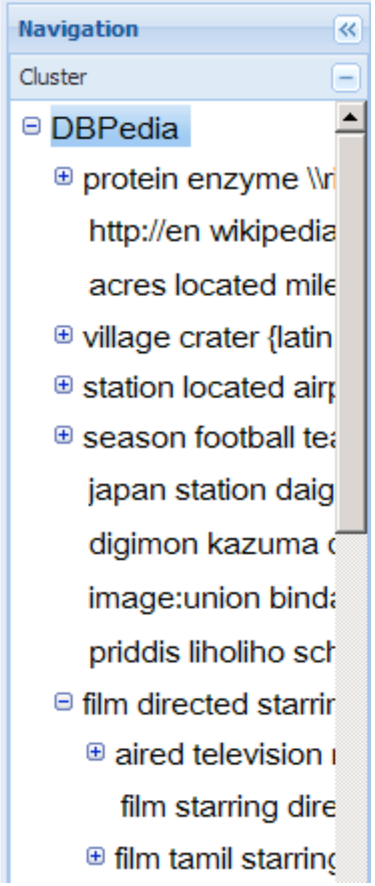
- Heterogeneity

- Consider a height predicate
  - Average value is 30 (Feet? Inches?)
  - But there are heights of buildings (in feet) and plants (in inches)
  - Average height of a building is 64 feet
  - Average height of a plant is 4 inches

- Prerequisite for meaningful profiling
- Volume of the data

ProLOD - Data Partitioning - Clustering

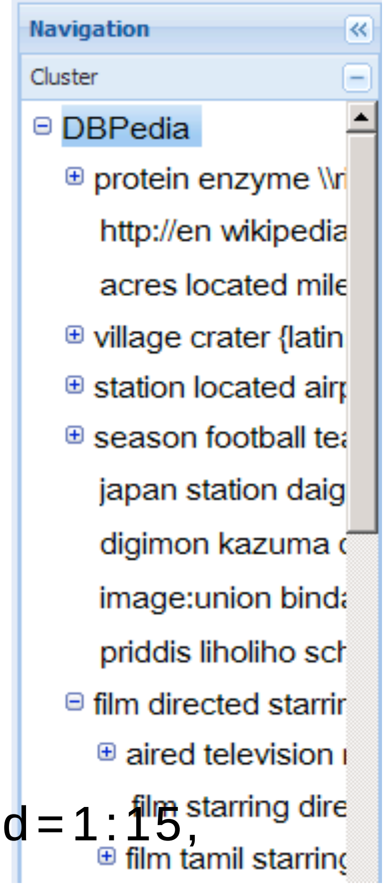
- Similarity of data entities
  - □ Schema Similarity = Jaccard Similarity
- Dissimilarity of data entities
  - □ Schema Dissimilarity = 1 – Schema Similarity
- Intra-Cluster Dissimilarity
  - □ Average pairwise Schema Dissimilarity
- Cluster Centroid
  - □ Schema of a cluster = Mean Schema
  - □ Threshold Mean Schema
    = Predicates required to be in t% of subjects
  - □ Top N Mean Schema (default)
    = N most frequent properties (N avg number of properties)



- ⊟ DBPedia
  - ⊞ protein enzyme \\r
  - http://en wikipedia
  - acres located mile
  - ⊞ village crater {latin
  - ⊞ station located airp
  - ⊞ season football tea
  - japan station daig
  - digimon kazuma c
  - image:union binda
  - priddis liholiho sch
  - ⊟ film directed starrir
    - ⊞ aired television
    - film starring dire
    - ⊞ film tamil starring

ProLOD - Data Partitioning - Clustering

- Iterative
  - □ Cluster data with k=2
  - □ While Cluster dissimilarity > threshold
    - ◇ Choose single Cluster C
    - ◇ Cluster C with k=2 (overall k increases)
- Hierarchical
  - □ recursive call of iterative K-Means
  - □ Predefined set of parameters to stop recursion
    - ◇ Max depth: 3
    - ◇ Max number of clusters in depth d: d=0:50, d=1:15, d=2:7
    - ◇ Max Cluster Dissimilarity: 0.3
    - ◇ Min Cluster Size: 100



⊟ DBPedia
- ⊞ protein enzyme \\r
- http://en wikipedia
- acres located mile
- ⊞ village crater {latin
- ⊞ station located airp
- ⊞ season football tea
- japan station daig
- digimon kazuma c
- image:union binda
- priddis liholiho sch
- ⊟ film directed starrir
  - ⊞ aired television
  - film starring dire
  - ⊞ film tamil starring

ProLOD - Data Partitioning - Labeling

- Use of textual subject descriptions
  - □ `rdf:comment`
  - □ `rdf:about`
  - □ `shortAbstract` (in DBpedia)

- Top k tf-idf weighted terms (default k=3, cluster is a document)
- Evaluation:
  - □ Given a grouping by `wikiUsesTemplate`
  - □ >56% of labels contain token from template name
  - □ More textual descriptions per cluster → higher percentage

- Top k predicates from Mean Schema

ProLOD - Schema Discovery

- Enables initial understanding of the actual structure of the data (set of triples does not expose much structural information)
- Determining the actual schema (e.g., distinct attributes of a cluster)
- Finding equivalent attributes (e.g., name, family name, and surname)
- Discovering poor attributes (i.e., those that do not contain useful values for most data entries)
- Discover attribute correlations
  - association rules
  - inverse relations
  - foreign key relationships

ProLOD - Schema Discovery

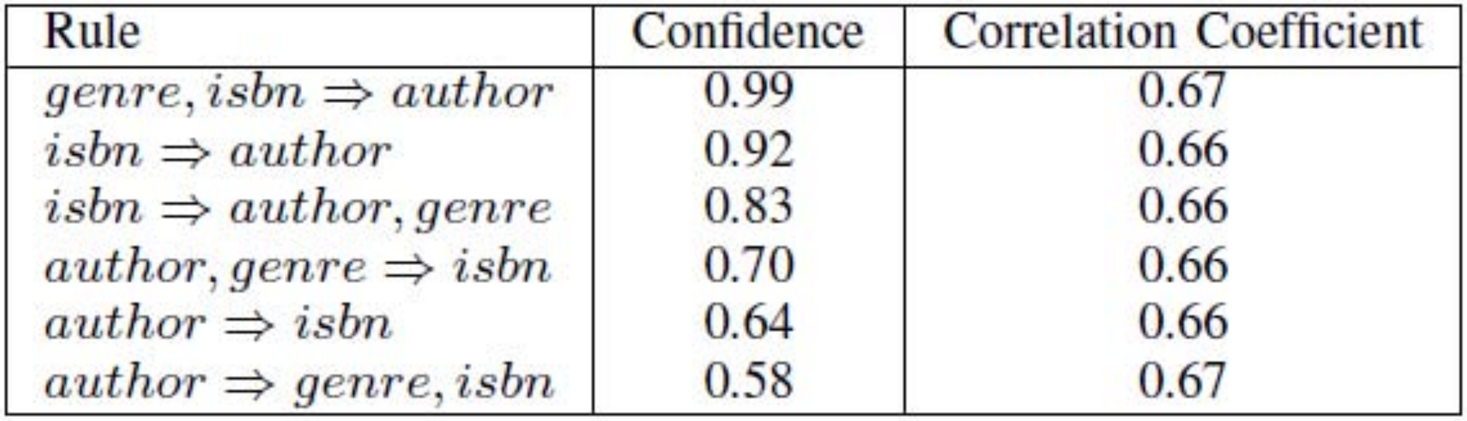
- Heterogeneity

- Determine set of attributes with ‘clean’ semantics from initial predicates
- Example: media cluster where entities have different predicates
  Consider author and/or developer predicates

- Most entities have author and developer, *distinct* semantics
  → Data ok, Clustering ok
- Most entities have either author or developer, *distinct* semantics
  → Data ok, Clustering questionable
- Most entities have author and/or developer, *similar* semantics
  → Data dirty, Clustering ok

ProLOD - Schema Discovery - Positive Association Rules

- Apriori Algorithm, Agrawal and Srikant, 1994
  - □ media cluster example:



Rule	Confidence	Correlation Coefficient
$genre, isbn \Rightarrow author$	0.99	0.67
$isbn \Rightarrow author$	0.92	0.66
$isbn \Rightarrow author, genre$	0.83	0.66
$author, genre \Rightarrow isbn$	0.70	0.66
$author \Rightarrow isbn$	0.64	0.66
$author \Rightarrow genre, isbn$	0.58	0.67

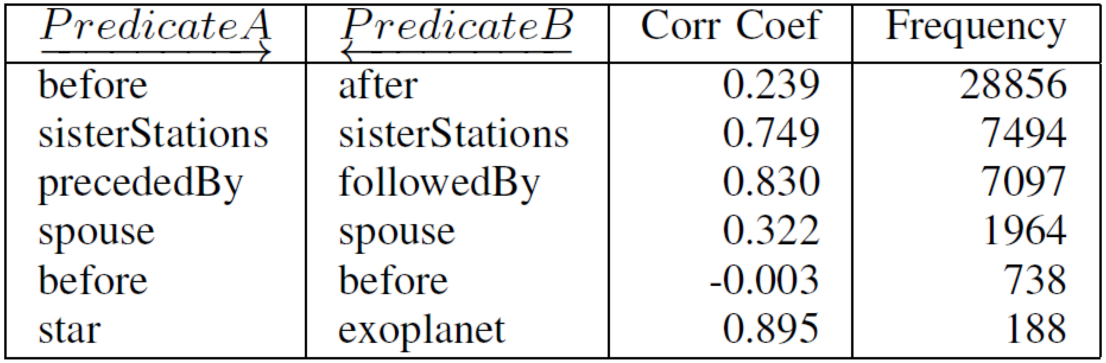
- Conclusion:
  - □ genre, isbn, author together form part of an entity’s schema
  - ➡ Assumption: complement each other
  - □ distinct semantics

ProLOD - Schema Discovery - Negative Association Rules

- Use of Correlation Coefficient, Antonie and Zaiane, 2004
- media cluster example:
  - name -> not(title)
- Conclusion:
  - Subjects from different domains in cluster → poorly built
    - Perform (sub)clustering with ProLOD
  - Semantic equivalence of predicates
    - Merge predicates in ProLOD

ProLOD - Schema Discovery - Inverse Links

- Subject X holds link to Subject Y via predicate $X \xrightarrow{A} Y$
- $X \xrightarrow{A} Y$ a$Y \xrightarrow{B} X$, then A and B are inverse links.
- Example:



$\underrightarrow{Predicate A}$	$\underleftarrow{Predicate B}$	Corr Coef	Frequency
before	after	0.239	28856
sisterStations	sisterStations	0.749	7494
precededBy	followedBy	0.830	7097
spouse	spouse	0.322	1964
before	before	-0.003	738
star	exoplanet	0.895	188

- Conclusion:
  - □ Redundancy of e.g. before/after and sisterStations
    - ◇ Fuse with ProLOD
  - □ Misuse of before
    - ◇ Exclude before with ProLOD

ProLOD - Statistics

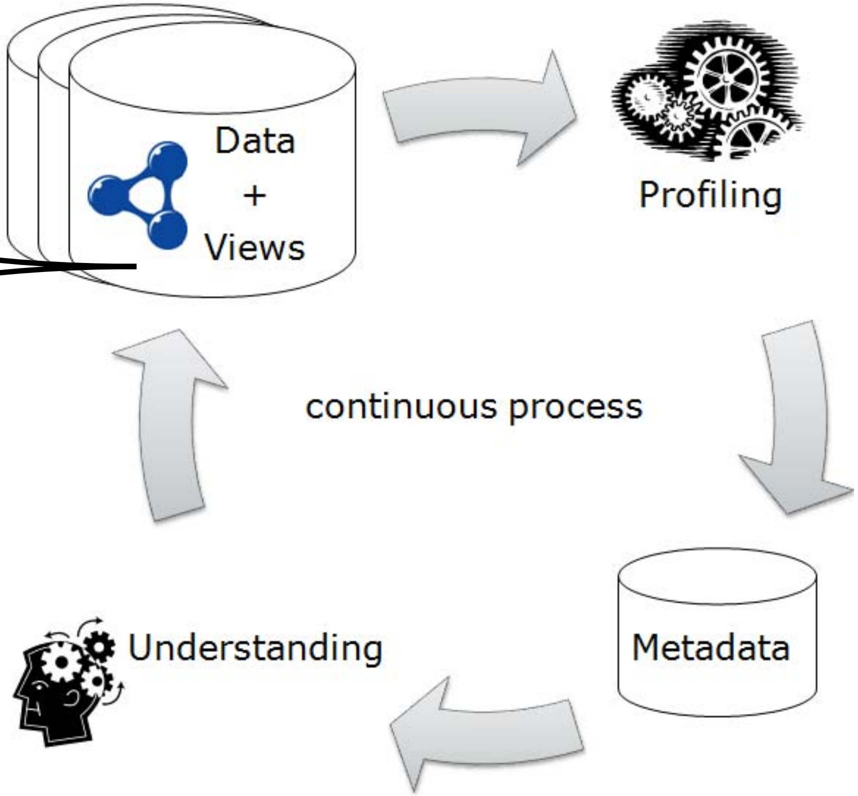
- (mostly) State-of-the-art Profiling for attribute values
- Distinction of values: literals, internal and external links

- Profiling for external links and literals
  - Data types
    (String, Text, Integer, Decimal, Date)
  - String → determine (normalized) patterns
  - Integers, Decimals → display value ranges
  - Set of user-defined keywords, and context rules
    - Months: Jan, Feb, Mar ...
      - Markus vs. Mar-06-2010 Aaaaaa vs. MONTH-99-9999
    - File extensions: .jpg, .mpg, ...
    - URL Schemas: http, ftp, ...

ProLOD - Usage

Clustering
+
Modifications
(merge, split ...)

Schema Discovery
+
Modifications
(filter, fuse, rename ...)



ProLOD Demo

http://youtu.be/_qyhVMOTbm0

Creating voiD Descriptions for Web-scale Data

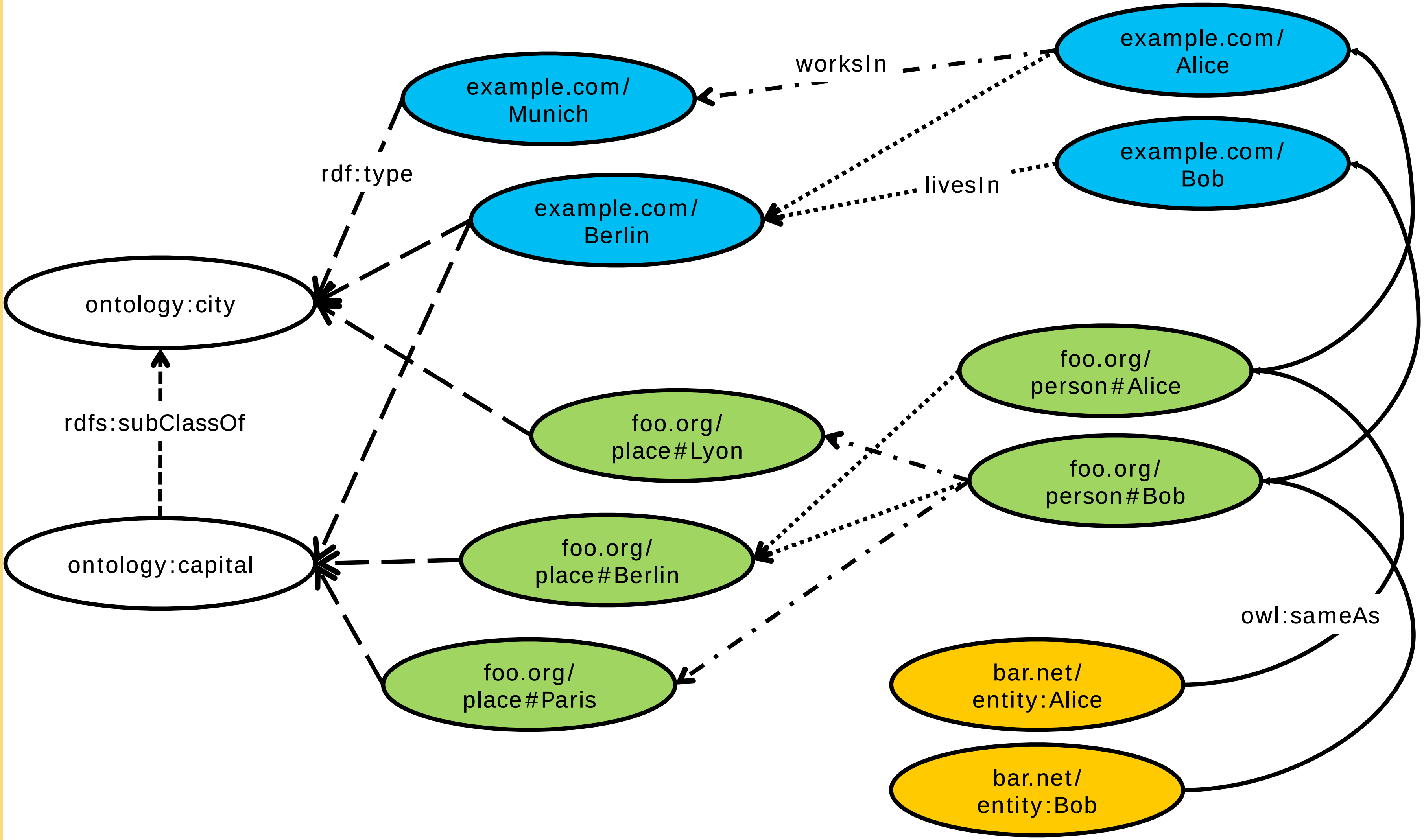
- Christoph Böhm, Johannes Lorey, Felix Naumann @ ISWC2010

- Scalable approach for segmenting, annotating, and enriching Linked Data sets
- Extend scope of voiD (Vocabulary of Interlinked Datasets)
  - Connected sets
    - 2 resources reside within the same connected dataset, iff there is a link of a specific type between them
  - Conceptual sets
    - 2 resources are contained in the same conceptual dataset, iff they are of the same or of similar type

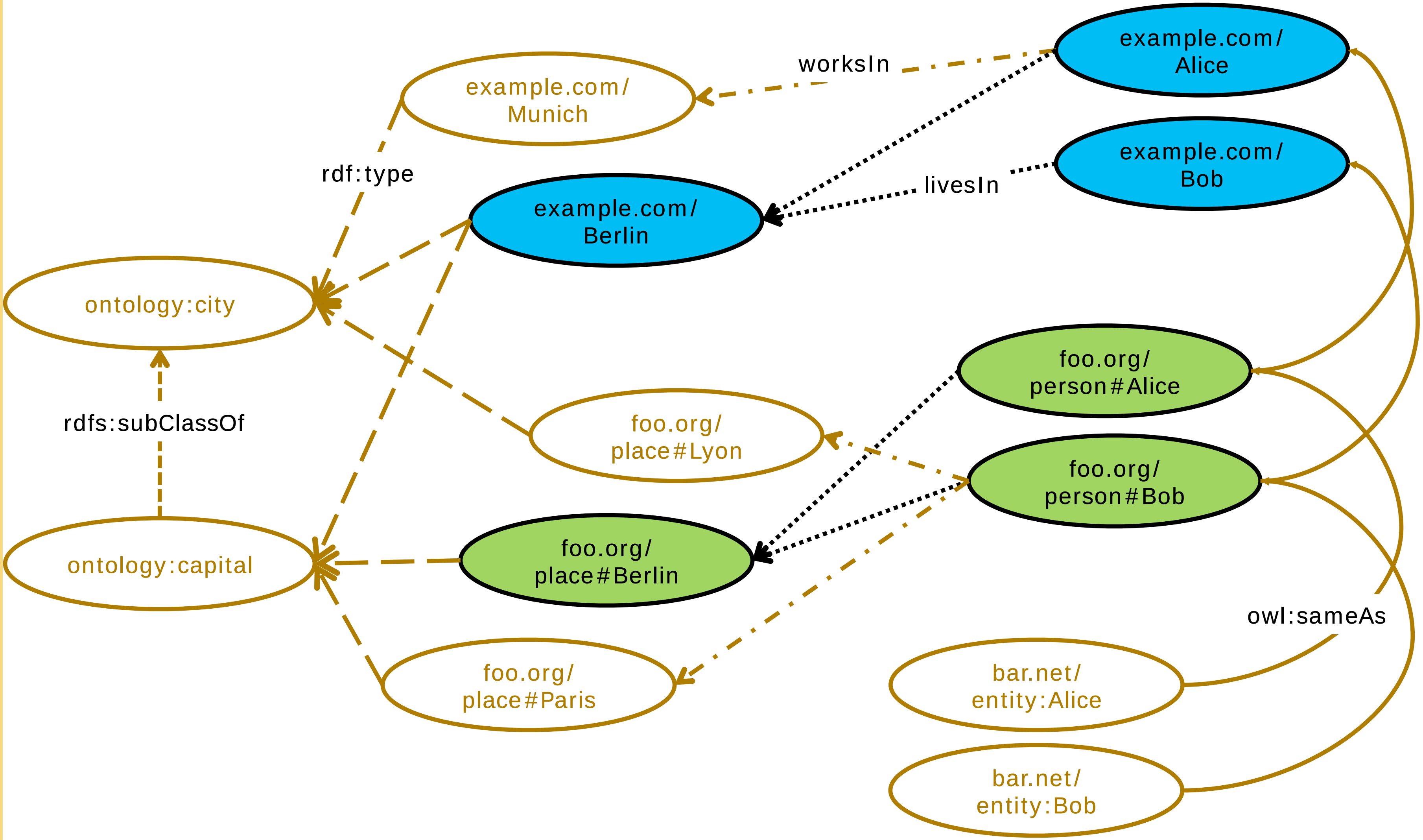
Creating voiD - Annotations per Dataset

- ✓ **void:datset**
- ✓ **void:linkset**
- ✓ **void:uriLookupEndpoint**
  - based on URI patterns of dataset resources
- ✓ **dcterms:description**
  - based on ranked list of subject types (rdf:type)
- ✓ **void:exampleResource**
  - based on dataset entity providing most statements
- ✓ **void:statItem**
  - various statistical information about dataset
- ✓ **void:vocabulary**
  - based on URIs of predicates

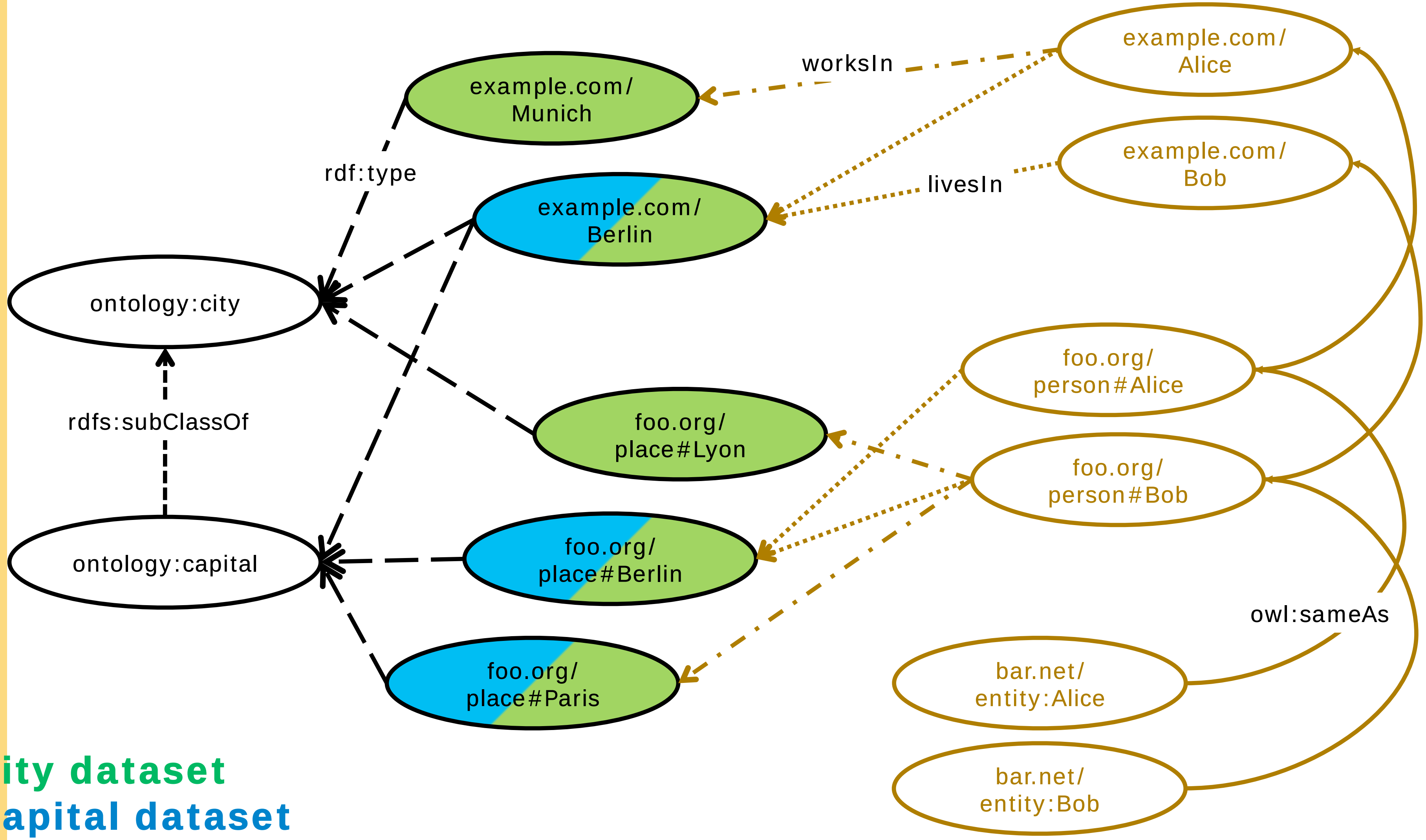
Creating voiD: Connectivity



Connected Datasets for voiD?



Conceptual Datasets for voiD?



city dataset
capital dataset

Identifying Relevant Sources for Data Linking using a Semantic Web Index

- Andriy Nikolov, Mathieu d’Aquin @ LDOW2011, WWW2011

- Two step approach:
  - Use subset of labels for keyword-based search on Semantic Web indexes to retrieve potentially relevant instances in external data sets
  - Use ontology matching techniques to filter out irrelevant sources by measuring semantic similarities between classes

Identifying Relevant Sources for Data Linking using a Semantic Web Index

- Keyword-based search for relevant instances:
  - Randomly select subset of individuals of belonging to a class (reduces number of search queries)
  - Query search engine (Sig.ma) for labels of each instance in subset
    - Sig.ma returns RDF document with references to instances, their sources and the classes they belong to
  - Aggregate search result
    - Load Sig.ma RDF documents in store and group instances by their sources
  - Data sets are ranked according to the numbers of returned instances

Identifying Relevant Sources for Data Linking using a Semantic Web Index

- Use ontology matching techniques to filter out irrelevant results
  - □ Use ontology matching algorithm (CIDER) to measure similarity between classes in original data sets and found classes
  - □ Filter out classes with low similarity index by applying a filter
  - □ Apply instance-based matching to BTC data set to map schemata based on ow:sameAs relations
  - □ Merge remaining classes with the classes obtained from the BTC schema mappings
  - □ Filter only instances that belong to the resulting class set

gold-miner

- Johanna Völkel, Mathias Niepert. http://code.google.com/p/gold-miner/

- Statistical schema induction
- Steps
  - □ Terminology acquisition from data set(s): classes and properties
  - □ Association rule mining
  - □ Ontology construction

Other existing approaches

- Conditional inclusion dependencies (Bauckmann, Naumann)
  - DBpedia person analysis in English and German DBpedia
  - Conditions on which German persons occur in English DBpedia

Conclusion

- Web of Data is growing
- Advantages:
  - Wealth of information
  - Easy, public access
  - Interesting domains
- Challenges:
  - Heterogeneity
    - Loose structure
    - Incomplete
    - Poorly formatted
    - Inconsistent
  - Volume of data

References

- Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, David Sonnabend. Profiling Linked Open Data with ProLOD. NTII2010, ICDE2010, 2010.
- Andriy Nikolov , Mathieu D'Aquin. Identifying Relevant Sources for Data Linking using a Semantic Web Index. LDOW2011, WWW2011, 2011.
- Andriy Nikolov, Enrico Motta. Capturing Emerging Relations between Schema Ontologies on the Web of Data. COLD2010, ISWC2010, 2010.
- Johanna Völkel, Mathias Niepert. Statistical Schema Induction. ESWC2011, 2011.