Datenrepräsentation der Wikipedia Enzyklopädie

Workshop

Berufetag 2016

Prof. Dr. Thomas Riechert

Hochschule für Technik, Wirtschaft und Kultur Leipzig
Fakultät Informatik, Mathematik und Naturwissenschaften

Anton-Philipp-Reclam Schule

Leipzig, 23.06.2016

Das Web verändert unsere Gesellschaft

- Soziale Kontakte (Social Networking-Plattformen, Blogs, ...)
- Economics (Kauf, Verkauf, Werbung, ...)
- Verwaltung (eGovernment)
- Work-Life (Informationsbeschaffung und-Sharing)
- Freizeit (Spiele, Rollenspiel, Kreativität, ...)
- Bildung (eLearning, Web als Informationssystem, ...)

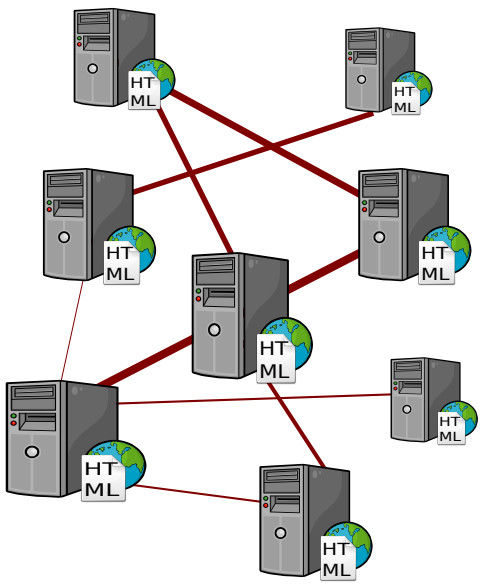
Grundzutaten für das Semantic Web

- Offene Standards für die Beschreibung von Informationen über das Web
- Verfahren zur Gewinnung weiterer Informationen aus solchen Beschreibungen

- Die Standards und deren Nutzung sind Kernelemente dieses Kurses.

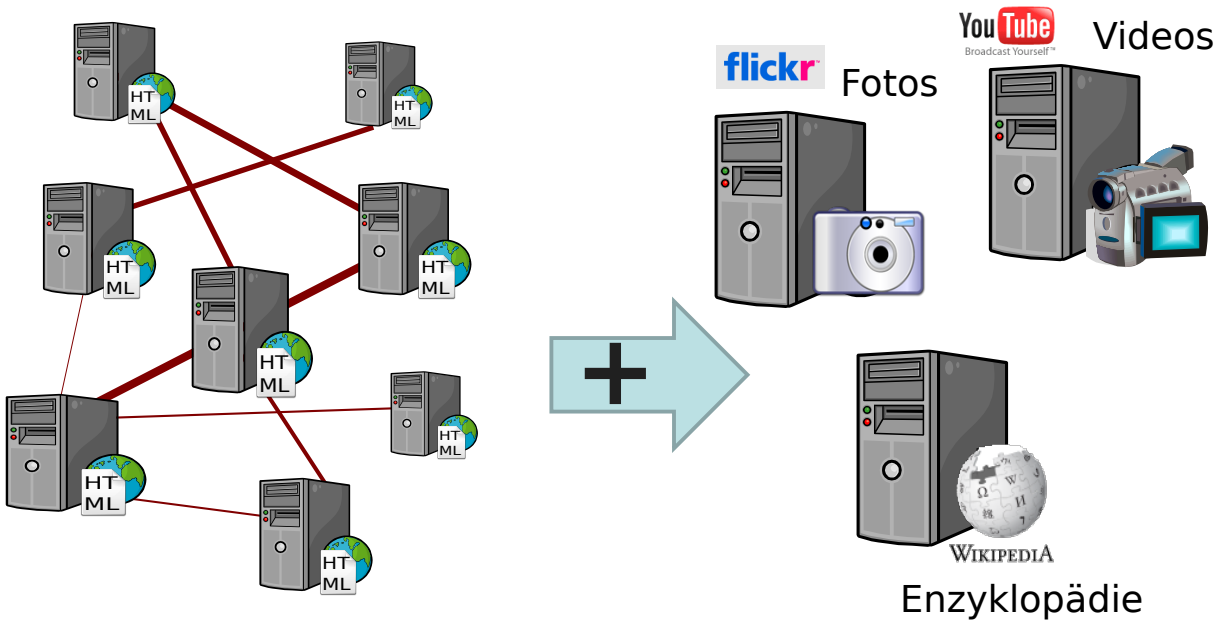
Web 1.0 - Das Netz der Hypertext-Dokumente (1992)

- Viele verschiedene Webseiten mit textuellem Inhalt



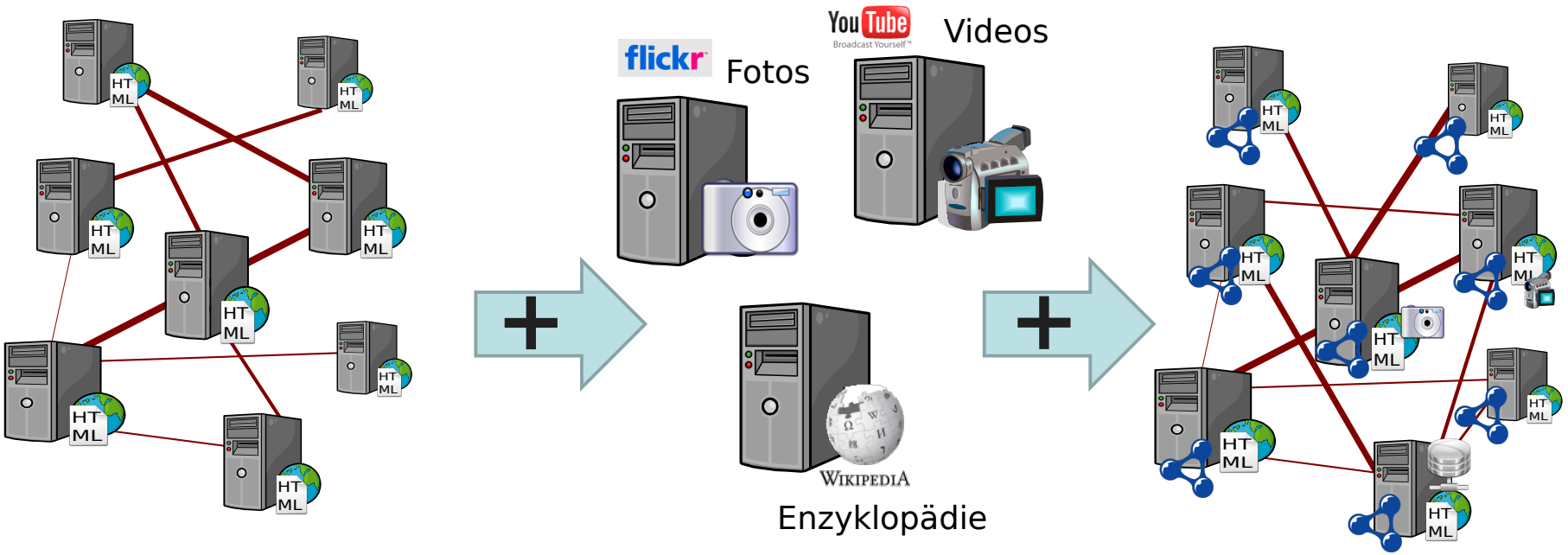
Web 2.0 - Das Web der Nutzerbeteiligung (2003)

- Viele verschiedene Webseiten mit textuellem Inhalt
- \+ Wenige sehr große Webseiten spezialisiert auf spezifische Inhaltstypen



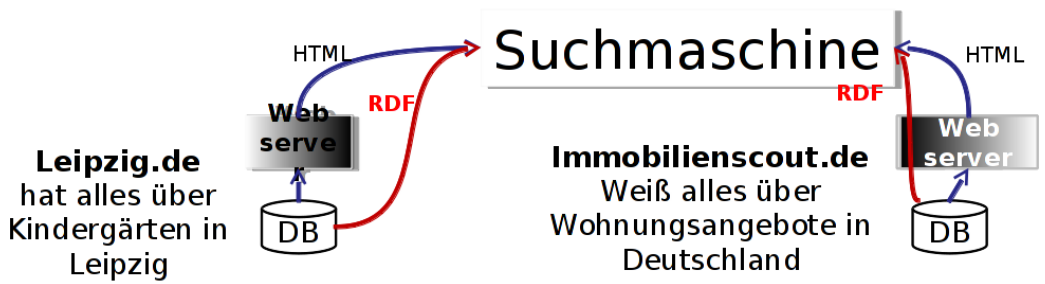
Web 3.0 - Das Web der Daten

- Viele verschiedene Webseiten mit textuellem Inhalt
- \+ Wenige sehr große Webseiten spezialisiert auf spezifische Inhaltstypen
- \+ Viele Webseiten enthalten und verweisen auf semantisch strukturierte Inhalte



Warum noch ein Web?

- Es gibt viele Suchanfragen welche von gängigen Suchmaschinen nicht erfüllt werden können:
  - Wohnungen mit gut bewerteter Miete, Thai Restaurants in der Nähe
  - Bi-linguale Englisch-Deutsch Kinderbetreuung in Berlin, erreichbar in 15 Minuten von Ihrem Arbeitsplatz
  - Kinder-freundliche Urlaubsziele mit Kultur-und Sportaktivitäten
  - Forscher in Süd-Ost Asien im Bereich Information Retrieval
  - ERP-Dienstleister mit Sitz in Wien und Berlin
- Wir haben unbewusst gelernt, Suchmaschinen solche Fragen nicht zu stellen.
- Im Prinzip ist das erforderliche Wissen im Web vorhanden.



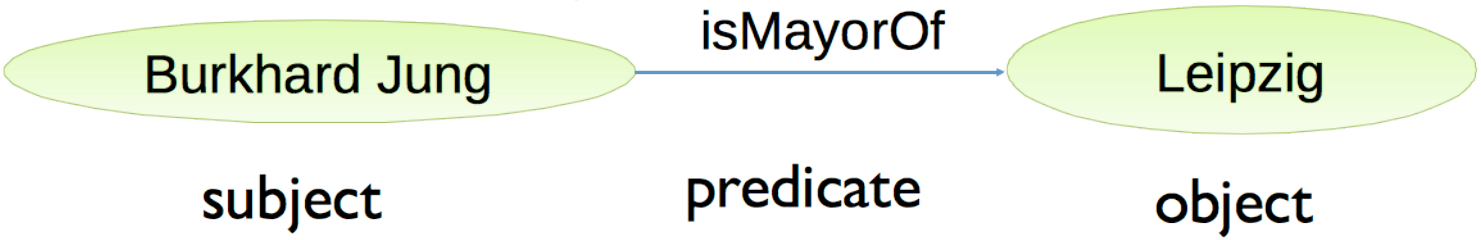
Exkurs: Syntax vs. Semantik

- **Syntax** (von grch. Zusammenstellung, Satzbau) steht für die (normative) Struktur von Daten, d.h. sie charakterisiert, was "wohlgeformte" Daten sind.
- **Semantik** (grch. zum Zeichen gehörend) steht für die Bedeutung von Daten, d.h. sie charakterisiert beispielsweise, welche inhaltliche Schlussfolgerungen sich ziehen lassen.

- `4+)=( syntaktisch falsch`
- `3+4=12 syntaktisch richtig -- semantisch falsch`
- `3+4=7 syntaktisch richtig -- semantisch richtig`

Resource Description Framework – RDF

Die Informationen werden in RDF in Tripel (auch Äußerungen, Fakten) dargestellt:

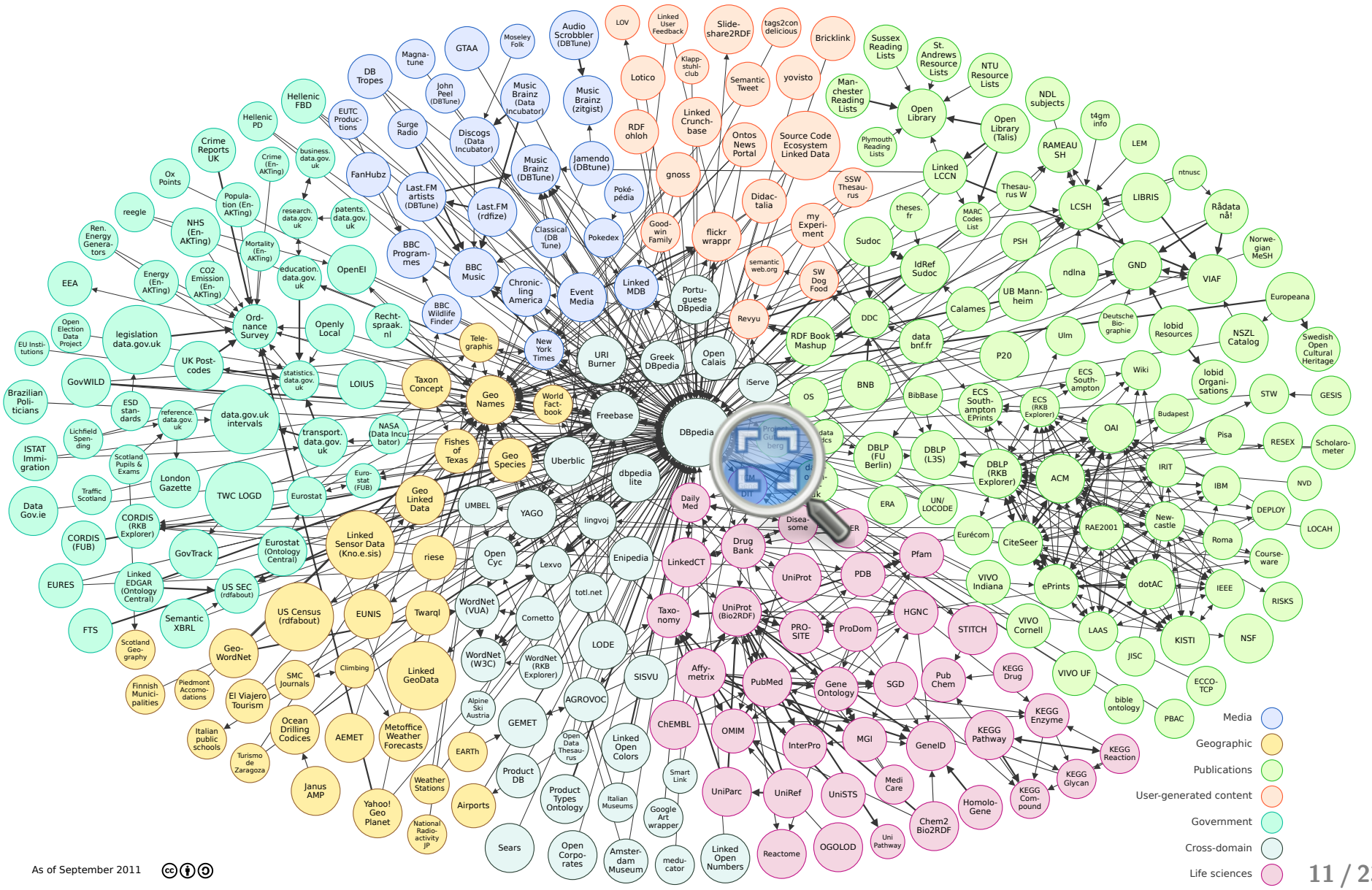


- Angelehnt an linguistische Kategorien, aber nicht immer konsistent
- Erlaubt Zuordnungen:
  - Subjekt: URI oder leerer Knoten
  - Prädikat: URI (Eigenschaft)
  - Objekt: URI, leerer Knoten oder Literal
- Knoten und Kantenbeschriftungen sollte eindeutig sein, so dass die ursprüngliche Grafik aus der Triple Liste rekonstruierbar ist

RDF Schema

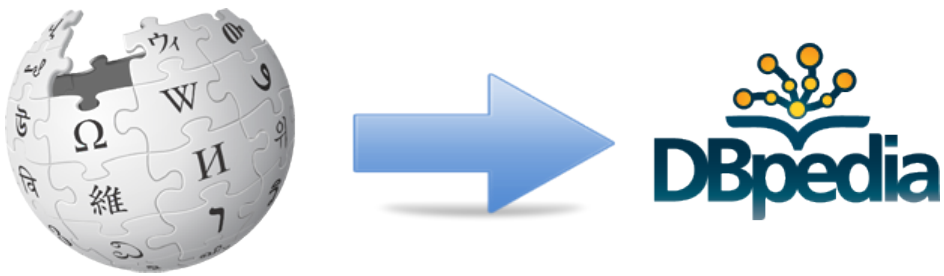
- Nicht alle Tripel sinnvoll:
  - Cinema
  - AlbertEinstein
  - 2012
- Wie können wir die Verwendung von RDF beschränken?
- RDF Schema erlaubt es, Klassen, Eigenschaften zu definieren und deren Nutzung einzuschränken.

Linking Open Data Cloud Diagramm



DBpedia: Wikidaten zu RDF-Daten Transformation

- Extrahiert strukturierte Daten von Wikipedia und macht diese als RDF-Daten verfügbar.
  - Ermöglicht komplexe Anfragen an Wikipedia
  - Verlinkung zu anderen Datensets im Web of Data
  - Repräsentiert einen Community-Konsens



- Semi-strukturiertes Wiki Markup $\rightarrow$ strukturierte Informationen
- Gemeinsames Ziel mit Wikidata aber anderer Ansatz
- DBpedia ist ein Community-Projekt, siehe `http://dbpedia.org` für eine vollständige Liste der Mitwirkenden

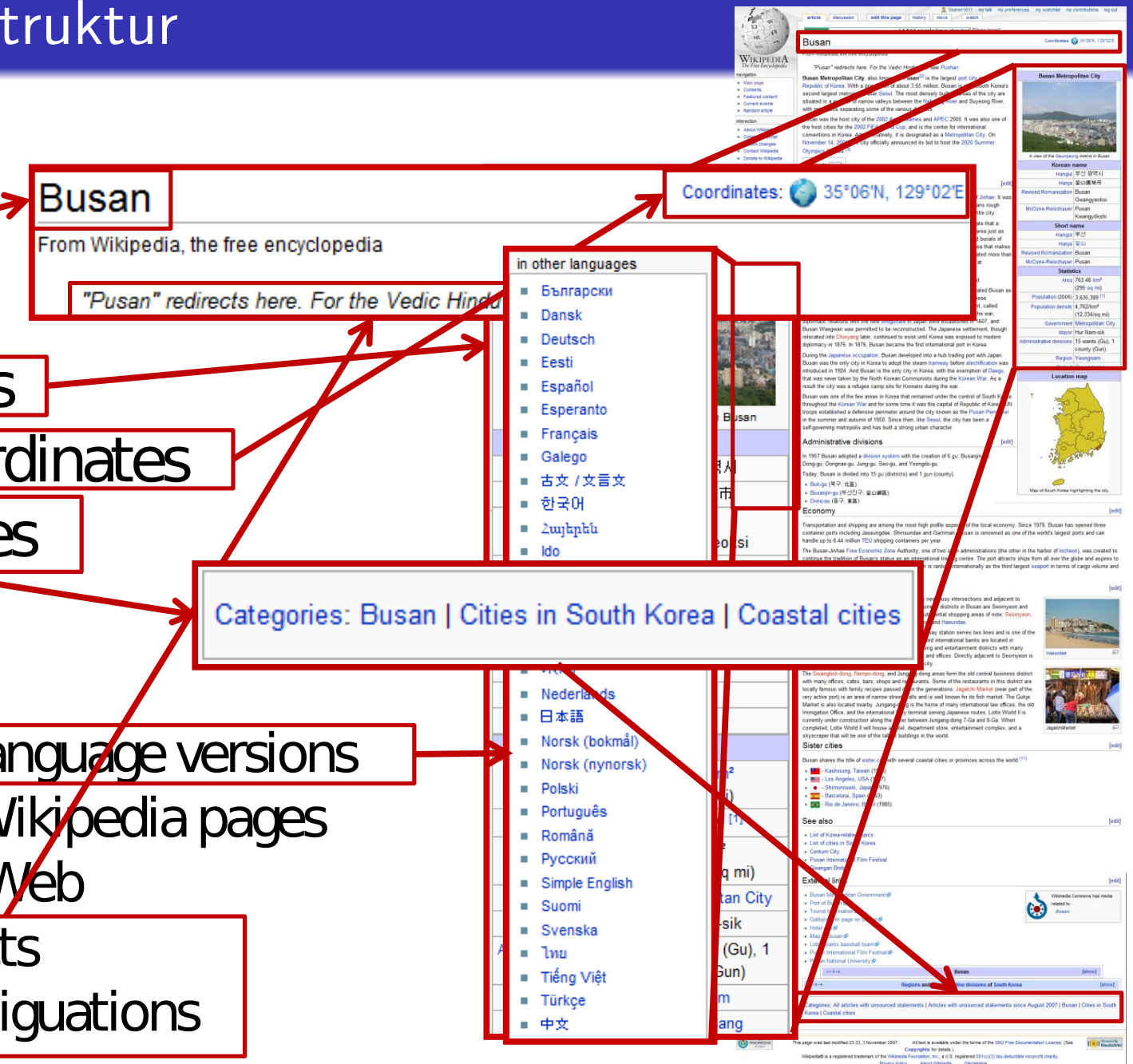
Wikipedia Einschränkungen

Einfache Fragen sind schwer mit Wikipedia zu beantworten:

- Was haben Innsbruck und Leipzig gemeinsam?
- Wie heißen die Bürgermeister der mitteleuropäischen Städte, welche höher als 1000m liegen?
- In welchen Filmen haben sowohl Brad Pitt als auch Angelina Jolie mitgespielt?
- Alle Fußballer, die als Torhüter für einen Verein spielen, welcher ein Stadion mit mehr als 40.000 Plätzen besitzt und die in einem Land, mit mehr als 10 Millionen Einwohnern, geboren sind

Wikipedia: Struktur

- Title
- Abstract
- Infoboxes
- Geo-coordinates
- Categories
- Images
- Links
  - other language versions
  - other Wikipedia pages
  - To the Web
  - Redirects
  - Disambiguations



Infoboxen - Templates

Wikitext-Syntax

```
{{Infobox Korean settlement
| title        = Busan Metropolitan City
| img          = Busan.jpg
| imgcaption   = A view of the [[Geumjeong]] district in Busan
| hangul       = 부산 광역시
...
| area_km2         = 763.46
| pop          = 3635389
| popyear      = 2006
| mayor        = Hur Nam-sik
| divs         = 15 wards (Gu), 1 county (Gun)
| region       = [[Yeongnam]]
| dialect      = [[Gyeongsang]]
}}
```

http://dbpedia.org/resource/Busan

RDF representation

```
dbp:Busan   dbpp:title     "Busan Metropolitan City"
dbp:Busan   dbpp:hangul    "부산 광역시"@Hang
dbp:Busan   dbpp:area_km2 "763.46"^xsd:float
dbp:Busan   dbpp:pop       "3635389"^xsd:int
dbp:Busan   dbpp:region    dbp:Yeongnam
dbp:Busan   dbpp:dialect   dbp:Gyeongsang
...
```

Busan Metropolitan City

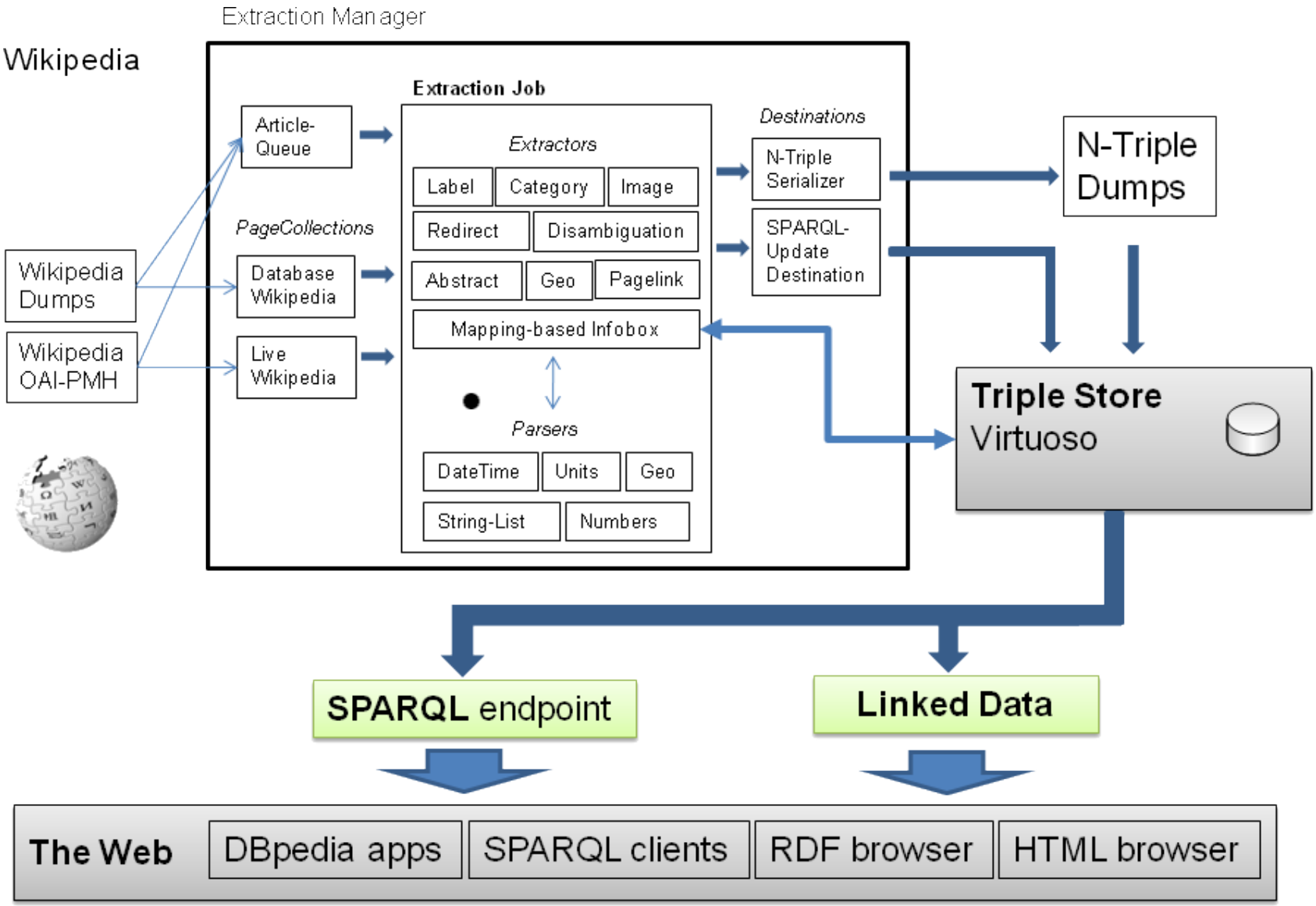
A view of the Geumjeong district in Busan

Korean name	
Hangul	부산 광역시
Hanja	釜山廣域市
Revised Romanization	Busan Gwangyeoksi
McCune-Reischauer	Pusan Kwangyŏkshi
Short name	
Hangul	부산
Hanja	釜山
Revised Romanization	Busan
McCune-Reischauer	Pusan
Statistics	
Area	763.46 km² (295 sq mi)
Population (2006)	3,635,389 [1]
Population density	4,762/km² (12,334/sq mi)
Government	Metropolitan City
Mayor	Hur Nam-sik
Administrative divisions	15 wards (Gu), 1 county (Gun)
Region	Yeongnam
Dialect	Gyeongsang

Resultierende Wissensdatenbank

- Beschreibungen von ca. 3,4 Millionen Dinge (1,5 Millionen in einer konsistenten Ontologie, einschließlich 312.000 Personen, 413.000 Plätze, 94.000 Musik-Alben, 49.000 Filmen, 15.000 Videospielen, 140.000 Organisationen, 146.000 Arten, 4.600 Erkrankungen
- Labels und Abstracts für diese 3,2 Millionen Dinge in bis zu 92 verschiedenen Sprachen; 1.460.000 Links zu Bildern und 5.543.000 Links zu externen Webseiten; 4.887.000 externe Links in andere RDF-Datensätze, 565.000 Wikipedia-Kategorien und 75.000 YAGO Kategorien
- insgesamt mehr als 1 Milliarde Fakten (das heißt RDF Tripel): 257M aus englischer Ausgabe, 766M von anderen Sprachausgaben

DBpedia Architektur



Wikipedia Fragen beantworten

Einfache Fragen sind schwer mit Wikipedia zu beantworten:

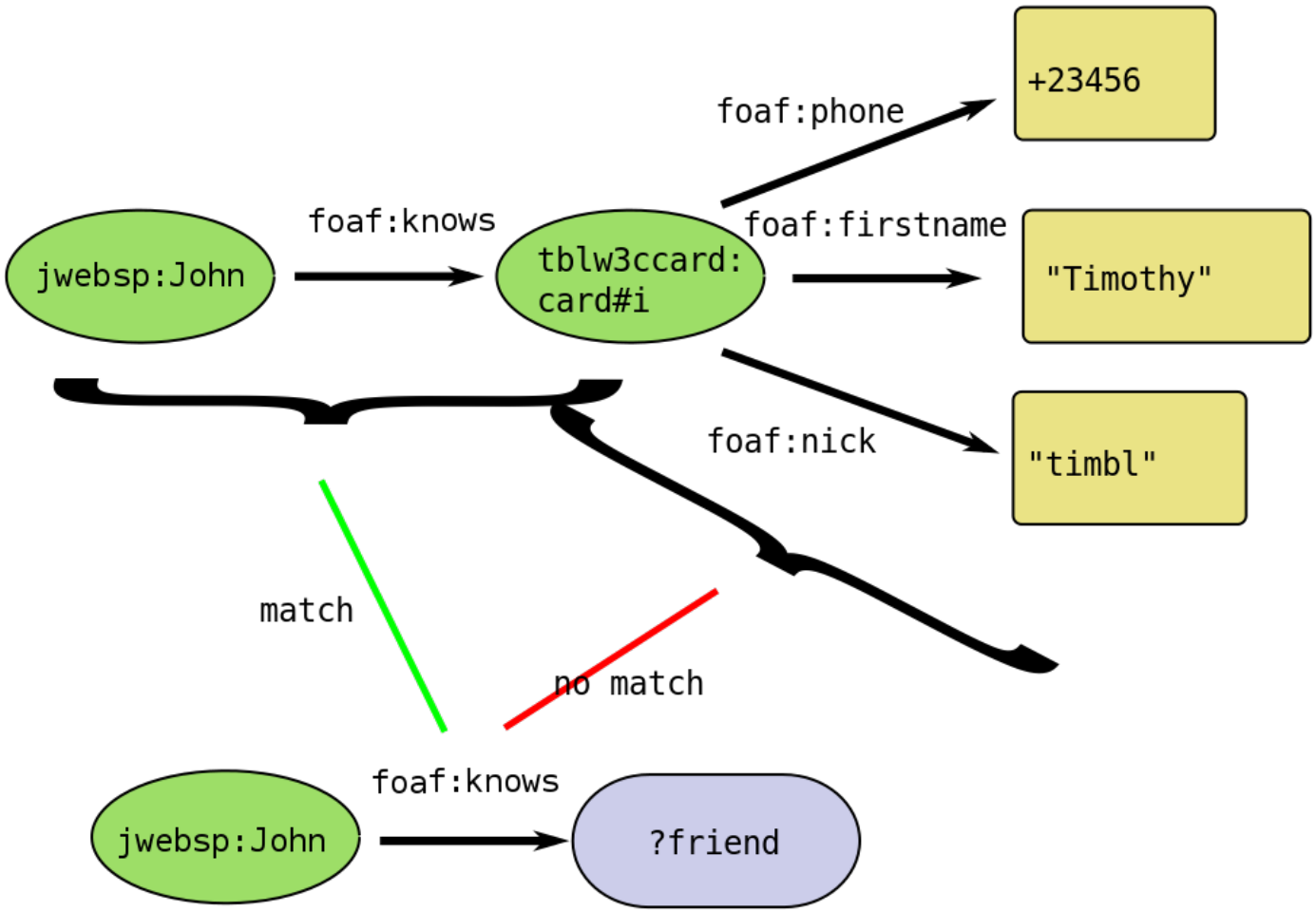
- Was haben Innsbruck und Leipzig gemeinsam?
- Wie heißen die Bürgermeister der mitteleuropäischen Städte, welche höher als 1000m liegen?
- In welchen Filmen haben sowohl Brad Pitt als auch Angelina Jolie mitgespielt?
- Alle Fußballer, die als Torhüter für einen Verein spielen, welcher ein Stadion mit mehr als 40.000 Plätzen besitzt und die in einem Land, mit mehr als 10 Millionen Einwohnern, geboren sind

DBpedia kann diese Fragen beantworten

- bietet einen öffentlichen SPARQL Endpunkt `http://dbpedia.org/sparql`
- hosted auf einem OpenLink Virtuoso Server

SPARQL – Query Language for RDF

SELECT * WHERE { jwebsp:John foaf:knows ?friend }



SPARQL-Endpunkt

```
SELECT ?name ?birth ?description ?person WHERE {
    ?person  dbp:birthPlace  dbp:Berlin .
    ?person  skos:subject    dbp:Cat:German_musicians .
    ?person  dbp:birth       ?birth .
    ?person  foaf:name       ?name .
    ?person  rdfs:comment    ?description .
  FILTER (LANG(?description) = 'en') .
} ORDER BY ?name
```

name	birth	description	person
"Moser, Edda"@de	"1938-10-27"^^xsd:date	"The German soprano Edda Moser was born on October 27, 1938 in Berlin, Germany. She is the daughter of the musicologist Hans Joachim Moser."@en	:Edda_Moser
"Möbius, Ralph Christian""@de	"1950-01-09"^^xsd:date	"Rio Reiser (January 9, 1950 - August 20, 1996), was a German rock musician and singer of the famous rock group Ton Steine Scherben. He was born Ralph Christian Möbius in Berlin and died at the age of 46 in the little German town of Fresenhagen. Rio Reiser was politically active during his whole life. In the early 70ies he participated in the squatter scene, for which he wrote the famous Rauchhaussong.""@en	:Rio_Reiser
"Straube, Karl"@de	"1873-01-06"^^xsd:date	"Montgomery Rufus Karl/Carl Siegfried Straube (January 6, 1873, Berlin - April 27, 1950, Leipzig) was a German church musician , organist, and choral conductor, famous above all for championing the abundant organ music of Max Reger."@en	:Karl_Straube
"Tricht, Käte van""@de	"1909-10-22"^^xsd:date	"Käte van Tricht (October 22, 1909–July 13, 1996), was a German organist, pianist, harpsichordist, and pedagogue.""@en	:K%C3%A4te_van_Tricht
"Urlaub, Farin"@de	"1963-10-27"^^xsd:date	"Jan Ulrich Max Vetter, better known as Farin Urlaub (like German "Fahr in Urlaub!" ("Go on holiday!"), after his love of travelling) was born on October 27, 1963 in what was then West Berlin, Germany. He is best known as the guitarist/vocalist for the German punk rock band Die Ärzte.""@en	:Farin_Urlaub
"Voormann, Klaus"@de	"1938-04-29"^^xsd:date	"Klaus Voormann (born 29 April 1938) is a German artist, musician, and record producer who was associated with the early days of The Beatles in Hamburg and later designed the cover of their album Revolver."@en	:Klaus_Voormann

URI / IRI Systeme

- http://{lang.}dbpedia.org ist die Haupt-Domain
- Für jeden Artikel gibt es eine DBpedia Ressource in der Form: http ://{lang.}dbpedia.org/resource/{Article Name}
- Eigenschaften aus dem Infobox Extraktor verwenden http://{lang.}dbpedia.org/property/{Namespace}

- Ontologie ist global für alle Sprachen und unter dem http://dbpedia.org/ontology/{Namespace}
- Hinweis: für die englische Sprache wird kein Sprach-Code verwendet

Beispiel-Ressourcen

- `http://rawgit.com/triechert/htwk_dbPediaRdfReader/develop/index.html`
  - Tool ist initial implementiert von Tabias Hahn. – Vielen Dank!
- Leipzig
  - Wikipedia-Seite: `http://de.wikipedia.org/wiki/Leipzig`
  - DBpedia Seite: `http://de.dbpedia.org/page/Leipzig`
  - DBpedia Ressource: `http://de.dbpedia.org/resource/Leipzig`

Literaturempfehlungen

- `http://rawgit.com/triechert/workshop/master/berufetag/index.html`
- Fabian M. Fürste: **Linked Open Library Data: bibliographische Daten und ihre Zugänglichkeit im Web der Daten.** B.I.T.Verlag, 2011.
- John Breslin, Alexandre Passant und Stefan Decker: **The Social Semantic Web.** Heidelberg: Springer-Verlag, 2009.
- Pascal Hitzler, Markus Krötzsch, York Sure und Sebastian Rudolph. **Semantic Web.** eXamen.press. Springer, 2007.
- Liyang Yu. **A Developer's Guide to the Semantic Web.** Springer, 2011.