DBpedia Overview

Jens Lehmann
AKSW Research Group
Institute of Computer Science

UNIVERSITÄT LEIPZIG

Transforming Wikipedia into a Knowledge base

- ☺ Wikipedia is the 7th most popular website (according to Alexa.com)
- ☺ Maybe the finest example of truly collaboratively created content (>8M articles, >200 languages, >300.000 authors)
- ☺ Covers many possible topics and domains, articles are a result of a "community consensus"
- ⊖ Many inconsistencies can be found on different pages/language versions

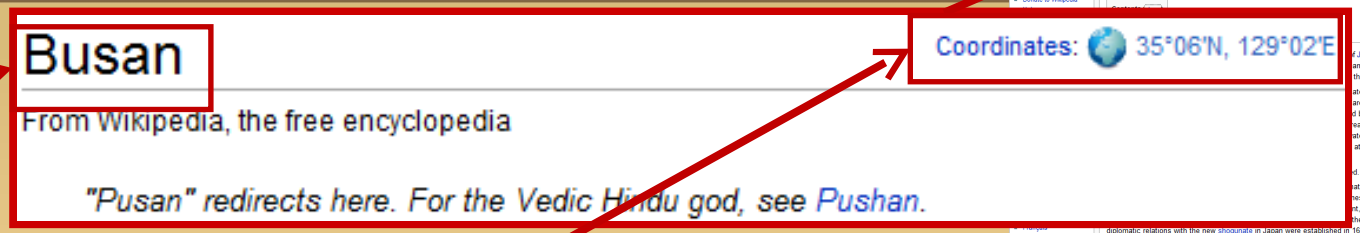
- ⊖ Not very well integrated with other data sources
- ⊖ Lacks structured representations of content which facilitate querying and search

Transforming Wikipedia into a Knowledge base

- Simple Questions – hard to answer:
  - What have Innsbruck and Leipzig in common**?**
  - Who are mayors of central European towns elevated more than 1000m**?**
  - Which films are longer than 4 hours and had a budget of less than $1 Million**?**
- The information required to answer these is contained in Wikipedia**!**
- How can we reveal structure and semantics of Wikipedia content?

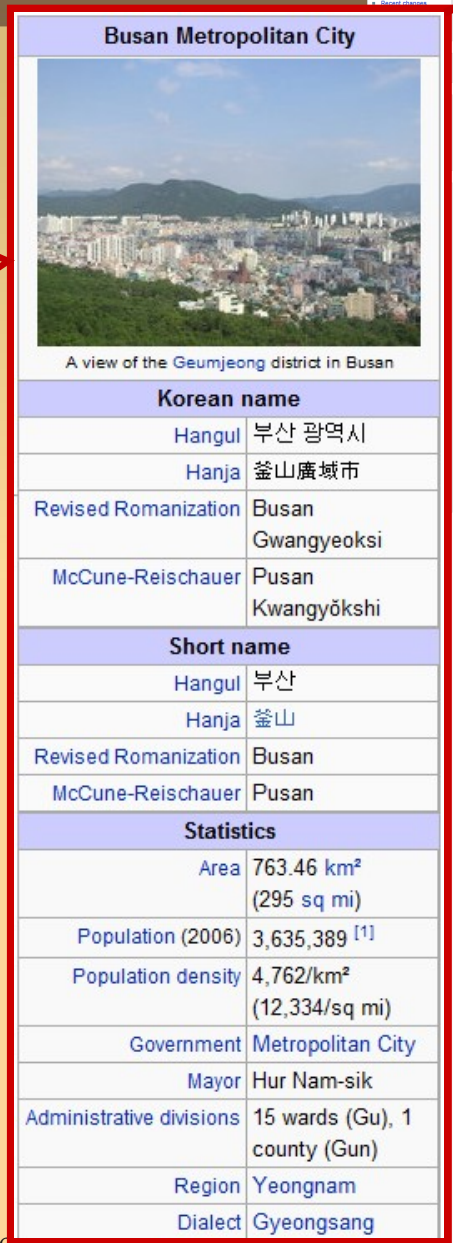
Structure in Wikipedia

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



Structure in Wikipedia

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



Busan Metropolitan City

A view of the Geumjeong district in Busan

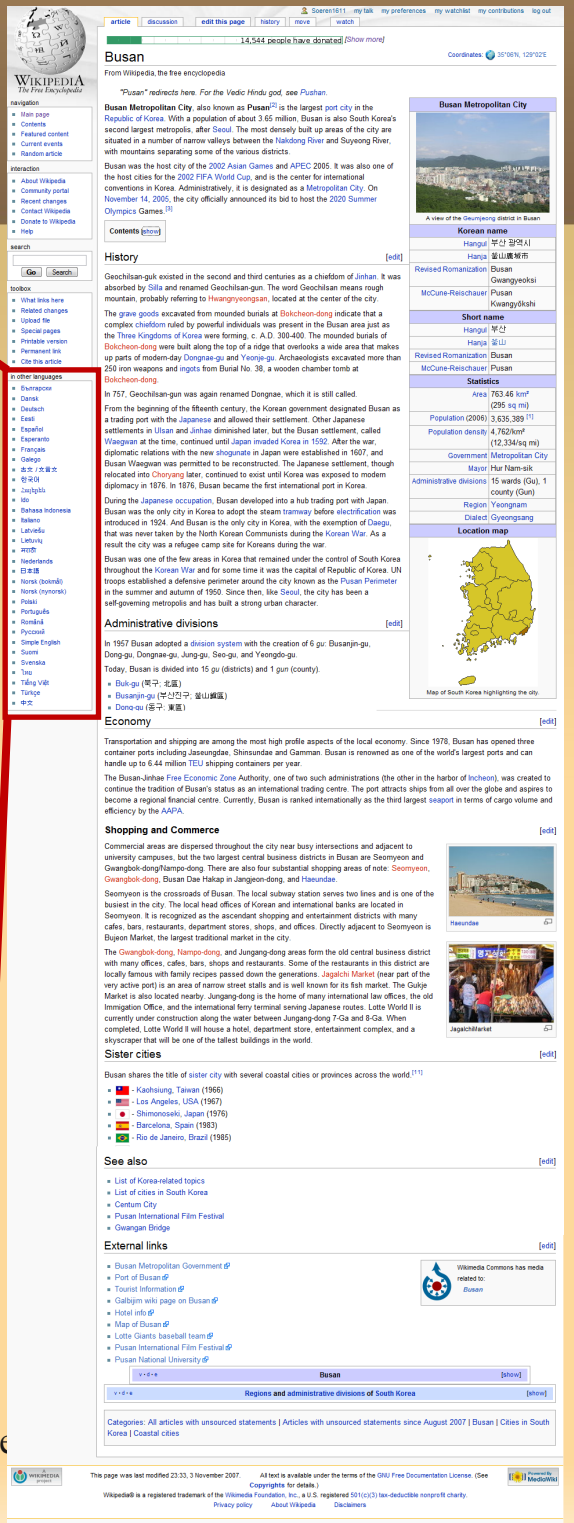
Korean name	
Hangul	부산 광역시
Hanja	釜山廣域市
Revised Romanization	Busan Gwangyeoksi
McCune-Reischauer	Pusan Kwangyŏkshi
Short name	
Hangul	부산
Hanja	釜山
Revised Romanization	Busan
McCune-Reischauer	Pusan
Statistics	
Area	763.46 km² (295 sq mi)
Population (2006)	3,635,389 [1]
Population density	4,762/km² (12,334/sq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

Structure in Wikipedia

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

in other languages

- Български
- Dansk
- Deutsch
- Eesti
- Español
- Esperanto
- Français
- Galego
- 古文 / 文言文
- 한국어
- Հայերեն
- Ido
- Bahasa Indonesia
- Italiano
- Latviešu
- Lietuvių
- मराठी
- Nederlands
- 日本語
- Norsk (bokmål)
- Norsk (nynorsk)
- Polski
- Português
- Română
- Русский
- Simple English
- Suomi
- Svenska
- ไทย
- Tiếng Việt
- Türkçe
- 中文



Structure in Wikipedia

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

Categories: Busan | Cities in South Korea | Coastal cities

Infobox 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]]
}}
```

RDF representation

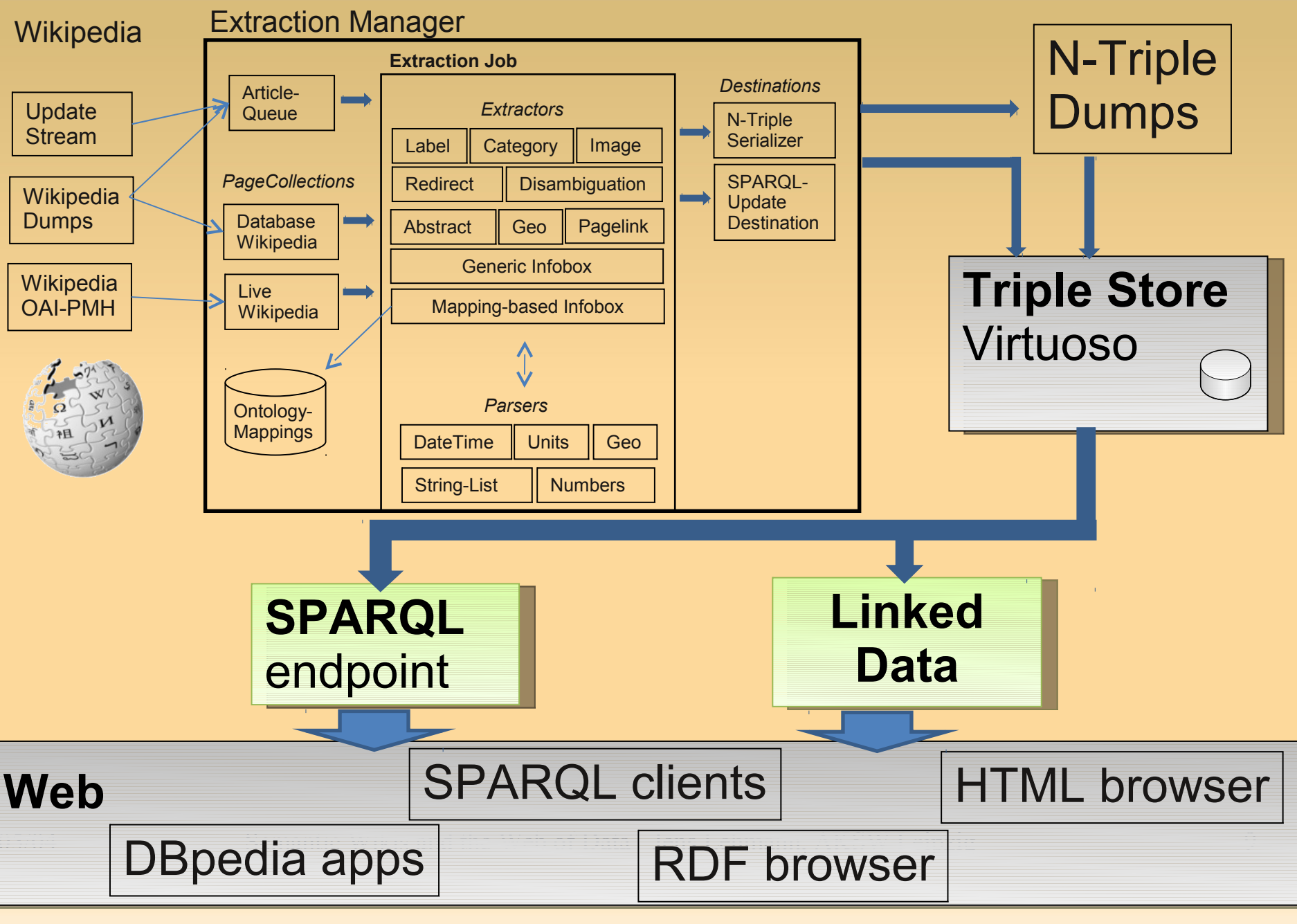
```
dbp:Busan   dbp:title     "Busan Metropolitan City"
dbp:Busan   dbp:hangul    "부산 광역시"@Hang
dbp:Busan   dbp:area_km2  "763.46"^^xsd:float
dbp:Busan   dbp:pop       "3635389"^^xsd:int
dbp:Busan   dbp:region    dbp:Yeongnam
dbp:Busan   dbp:dialect   dbp:Gyeongsang
...
```

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

DBpedia Architecture



Results

- Extraction in 97 languages
- DBpedia describes 3.5 million things, including at least 364,000 persons, 462,000 places, 99,000 music albums, 54,000 films, 148,000 organisations
- 672 million RDF triples
- Manual tests reveal: ~87% accuracy, (9% redundant information, 2% irrelevant, 1% errors)
- Very large multi-domain ontology

Dataset (En only)	Triples
Titles	2.7M
Abstracts	7.6M
External Links	3.2M
Categories	7.3M
Infoboxes (generic)	26.0M
Infoboxes (mapped)	7.0M
Yago Classes	2M
Geo-coordinates	450k
Properties	66k
Mapping to Flickr, DBLP, Eurostat, CIA-Factbook, Musicbrainz, Project Gutenberg, US Census, …	2.5M
Mapping to OpenCyc	45k

Hierarchies

- DBpedia Ontology Schema:
  - manually created for DBpedia (infoboxes)
  - 275 classes + 1335 properties; 20mio triples
- YAGO:
  - large hierarchy linking Wikipedia leaf categories to WordNet
  - 250,000 classes
- UMBEL (Upper Mapping and Binding Exchange Layer):
  - 20000 classes derived from OpenCyc
- Wikipedia Categories:
  - Not a class hierarchy (e.g. cycles), represented using SKOS
  - 415,000+ categories

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?

DBpedia SPARQL Endpoint

```
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

DBpedia Applications

- DBpedia Mobile: location aware mobile client for DBpedia
  - Uses current location and DBpedia to display map
  - Can navigate into other knowledge bases
- DBpedia Query Builder: user front end for building queries
- DBpedia Relationship Finder finds relation between two objects in DBpedia

DBpedia Applications (3rd party)

- Muddy Boots (BBC): Annotate actors in BBC News with DBpedia identifiers
- Open Calais (Reuters): named entity recognition; entities are connected via owl:sameAs to DBpedia, Freebase, Geonames
- Faviki: Social Bookmarking Tool uses DBpedia in backend to group tags etc. and multi-language support
- Topbraid Composer: ontology editor, which links entities to DBpedia based on their labels

The End

Thanks for your attention!