DBpedia – Live

Kollaborative Ontologieerstellung
durch die Wikipedia/DBpedia-Gemeinschaft

Sebastian Hellmann
AKSW

UNIVERSITÄT LEIPZIG

Gliederung

- Wikipedia

- DBpedia – Überblick (AKSW, FU Berlin, OpenLink)

- Kollaborative Ontologieerstellung

- Live Extraktion

Wikipedia

Im Spiegel am 19.04.2010:

"Wikipedia ist die größte Enzyklopädie der Welt, bald könnte sie die einzige verbliebene sein.

Der gedruckte Brockhaus, Microsofts CD-Rom-Lexikon Encarta sind eingestellt, die Encyclopædia Britannica hat wirtschaftliche Probleme. [...]

Diese Aufgabe ist an das Kollektiv der Internetnutzer übergegangen."

Wikipedia

- **Nature am 15.12.2005:**

  Vergleich von 42 wissenschaftlichen Artikeln

  Wikipedia ca. 4 Fehler pro Artikel

  Encyclopædia Britannica ca. 3 Fehler pro Artikel

  Fehler in Wikipedia wurden zeitig behoben.

  [Jim Giles, Internet encyclopaedias go head to head Nature 438, 900-901 (15 December 2005)]

Wikipedia

London 7.7.2005, 8:50 Uhr

Bombenanschlag auf die Londoner U-bahn

- Erster Eintrag in Wikipedia 18 Minuten später

- 2500 Benutzer erstellen innerhalb von 12 Stunden einen 14-seitigen Artikel, der alle anderen herkömmlichen Medienberichte im Detailgrad weit übertrifft.

[Tapscott & Williams, Wikinomics]

Wikipedia

- hochaktuell

- wenige Fehler, die schnell korrigiert werden können

- DAS Wissenskompendium der Menschheit

Wie kann das in Wikipedia vorhandene Wissen aufgeschlüsselt und nutzbar gemacht werden?

Wikipedia als Wissensbasis

- Einfache Fragen:

  - Bei welche Filmen von Tarantino hat Bruce Willis mitgespielt?

  - Wie heissen die Bürgermeister von Europäischen Städten mit einer Einwohnerzahl zwischen 20.000 und 100.000 Einwohnern?

  - Welche Fussballspieler haben als Torwart für einen Verein gespielt, der ein Stadium mit mehr als 40.000 Plätzen hat und wurden in einem Land geboren, dass mehr als 10 Millionen Einwohner hat?

- **Die Antworten sind in Wikipedia vorhanden!**

Wissen in Wikipedia

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

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]]
}}
```
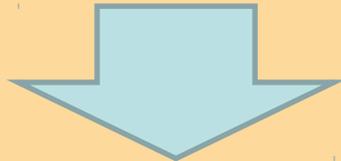
RDF representation

```
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	

04.05.2010

DBpedia, AKSW, LSWT

Datenset

- Extraktion in 92 Sprachen

- DBpedia enthält ca. 3,4 Millionen Dinge, mehr als 312,000 Personen, mehr als 413,000 Orte,...

- ca. 300 Millionen RDF Tripel

- Redirects ermöglichen das einfache Erstellen von Synonymlisten

- Links zwischen verschiedenen Sprachversionen ermöglichen Übersetzung von Begriffen

Dataset (En only)	Triples
Titles	3.4M
Abstracts	3.2M
External Links	5.5M
Categories	0.5M
Infoboxes (generic)	26.0M
Infoboxes (mapped)	7.0M
Yago Classes	75K
Geo-coordinates	1.5M
Properties	66k
Mapping to Flickr, DBLP, Eurostat, CIA-Factbook, Musicbrainz, Project Gutenberg, US Census, …	2.5M
Mapping to OpenCyc	45k

Kollaborative Ontologieerstellung

- **"A little Semantics goes a long way"** - Jim Hendler

Kollaborative Ontologieerstellung

- **"A little Semantics goes a long way"** - Jim Hendler
  - Expressive Abfragen können bereits mit der direkten Konvertierung beantwortet werden
  - http://tinyurl.com/2uhuow9

Kollaborative Ontologieerstellung

- **"A little Semantics goes a long way"** - Jim Hendler
  - Expressive Abfragen können bereits mit der direkten Konvertierung beantwortet werden
  - http://tinyurl.com/2uhuow9

- **More Semantics go a longer way...**
  - Kollaboratives Mapping und Schema erhöhen die Datenqualität
  - Mehr Treffer bei Abfragen

A closer look at infoboxes

George A. Romero

Romero in Venice, 2009

Born	George Andrew Romero February 4, 1940 (age 69) New York, NY, U.S.
Occupation	film director screenwriter editor actor
Spouse(s)	Christine Forrest (1981–present)

04.05.2010

Björk

Bjork performance at Rock en Seine, Paris 2007

Background information

Birth name	Björk Guðmundsdóttir
Born	21 November 1965 (age 43) Reykjavík, Iceland
Genres	Electronica, alternative rock, trip hop, IDM
Occupations	Musician, songwriter, producer, actress,

The Right Honourable
Gordon Brown
MP

Prime Minister of the United Kingdom

Incumbent

Assumed office
27 June 2007

Monarch	Elizabeth II
Preceded by	Tony Blair

Chancellor of the Exchequer

In office
2 May 1997 – 27 June 2007

Prime Minister	Tony Blair

A closer look at infoboxes

George A. Romero	
	Romero in Venice, 2009
Born	George Andrew Romero February 4, 1940 (age 69 New York, NY, U.S.
Occupation	film director screenwriter editor actor
Spouse(s)	Christine Forrest (1981–present)

04.05.2010

Björk	
	Bjork performance at Rock en Seine, Paris 2007
Background information	
Birth name	Björk Guðmundsdóttir
Born	21 November 1965 (age 43) Reykjavík, Iceland
Genres	Electronica, alternative rock, trip hop, IDM
Occupations	Musician, songwriter, producer, actress,

The Right Honourable **Gordon Brown** MP	
Prime Minister of the United Kingdom	
Incumbent	
Assumed office 27 June 2007	
Monarch	Elizabeth II
Preceded by	Tony Blair
Chancellor of the Exchequer	
In office 2 May 1997 – 27 June 2007	
Prime Minister	Tony Blair

A closer look at infoboxes

George A. Romero

Romero in Venice, 2009

Born	George Andrew Romero February 4, 1940 (age 69) New York, NY, U.S.
Occupation	film director screenwriter editor actor
Spouse(s)	Christine Forrest (1981– present)

01.05.2010

Björk

Bjork performance at Rock en Seine, Paris 2007

Background information

Birth name	Björk Guðmundsdóttir
Born	21 November 1965 (age 43) Reykjavík, Iceland
Genres	Electronica, alternative rock, trip hop, IDM
Occupations	Musician, songwriter, producer, actress,

The Right Honourable
Gordon Brown
MP

Prime Minister of the United Kingdom
Incumbent

Assumed office
27 June 2007

Monarch	Elizabeth II
Preceded by	Tony Blair

Chancellor of the Exchequer

In office
2 May 1997 – 27 June 2007

Prime Minister Tony Blair

```
{{Infobox musical artist
| Name          = Björk
| Img           = Björk Rock en Seine 2007 (2).jpg
| Img_capt      = Björk performance at Rock en Sei
| Img_size      =
| Landscape     =
| Background    = solo_singer
| Birth_name    = Björk Guðmundsdóttir
| Alias         =
| Born          = {{birth date and age|df=yes|196
| Died          =
| Origin        =
| Instrument    = [[Singer|Vocals]], [[Keyboard in                    ]], [[harp]]
|Voice_type     = [[Soprano]]<ref>http://www.spiri
| Genre         = [[Electronica]], [[alternative r
| Occupation    = [[Musician]], [[songwriter]], [[record producer|producer]], [[actor|actress]], [[fashion model]]
| Years_active  = 1977–present
| Label         = [[One Little Indian]], [[Elektra Records|Elektra]], [[Atlantic Records Group|Atlantic]], [[Polydor Records|Polydor]], [[Mother Re
[[Smekkleysa]]
| Associated_acts = [[The Sugarcubes]], [[Thom Yorke]], [[Tappi Tíkarrass]], [[KUKL (group)|KUKL]]
| URL           = [http://www.bjork.com/ www.bjork.com]
}}
```

Björk (Musician)
Occupation = Musician, Actor
Born = 21.12.1965, Reykjavík

```
{{Infobox Prime Minister
|honorific-prefix = <small>
|name             = Gordon B
|honorific-suffix = <br /><s
|image            = Gordon B
|imagesize        = 200px
|alt              = Gordon B
|office           = [[Prime
|monarch          = [[Elizab
|term_start       = 27 June
|term_end         =
|predecessor      = [[Tony Blair]]
|successor        =
|office2          = [[Chancellor of the Exchequer]]
|term_start2      = 2 May 1997
|term_end2        = 27 June 2007
|primeminister2   = [[Tony Blair]]
|predecessor2     = [[Kenneth Clarke]]
|successor2       = [[Alistair Darling]]
```

Brown (Prime Minister)
office = Prime Minister of the UK
birth_date = 20.4.1951
birth_place = Govan

```
{{Infobox actor
| image = George Romero, 66
| imagesize = 200px
| caption = Romero in Veni
| birthname = George Andre
| birthdate = {{birth date
| birthplace = [[New York,
| spouse = Christine Forres
| occupation = [[film dire
}}
```

Romero (Actor)
occupation = Actor, Editor
birthdate = 4.2.1940
birthplace = New York

04.05.2010

```
{{Infobox musical artist
| Name          = Björk
| Img           = Björk Rock en Seine 2007 (2).jpg
| Img_capt      = Björk performance at Rock en Sei
| Img_size      =
| Landscape     =
| Background    = solo_singer
| Birth_name    = Björk Guðmundsdóttir
| Alias         =
| Born          = {{birth date and age|df=yes|196
| Died          =
| Origin        =
| Instrument    = [[Singer|Vocals]], [[Keyboard in                    ]], [[harp]]
|Voice_type     = [[Soprano]]<ref>http://www.spiri
| Genre         = [[Electronica]], [[alternative r
| Occupation    = [[Musician]], [[songwriter]], [[record producer|producer]], [[actor|actress]], [[fashion model]]
| Years_active  = 1977–present
| Label         = [[One Little Indian]], [[Elektra Records|Elektra]], [[Atlantic Records Group|Atlantic]], [[Polydor Records|Polydor]], [[Mother Re
[[Smekkleysa]]
| Associated_acts = [[The Sugarcubes]], [[Thom Yorke]], [[Tappi Tíkarrass]], [[KUKL (group)|KUKL]]
| URL           = [http://www.bjork.com/ www.bjork.com]
}}
```

Björk (Musician)
Occupation = Musician, Actor
Born = 21.12.1965, Reykjavík

```
{{Infobox Prime Minister
|honorific-prefix = <small>
|name             = Gordon
|honorific-suffix = <br /><s
|image            = Gordon
|imagesize        = 200px
|alt              = Gordon
|office           = [[Prime
|monarch          = [[Elizab
|term_start       = 27 June
|term_end         =
|predecessor      = [[Tony Blair]]
|successor        =
|office2          = [[Chancellor of the Exchequer]]
|term_start2      = 2 May 1997
|term_end2        = 27 June 2007
|primeminister2   = [[Tony Blair]]
|predecessor2     = [[Kenneth Clarke]]
|successor2       = [[Alistair Darling]]
```

Brown (Prime Minister)
office = Prime Minister of the UK
birth_date = 20.4.1951
birth_place = Govan

```
{{Infobox actor
| image = George Romero, 66
| imagesize = 200px
| caption = Romero in Venic
| birthname = George Andrew
| birthdate = {{birth date
| birthplace = [[New York,
| spouse = Christine Forres
| occupation = [[film direc
}}
```

Romero (Actor)
occupation = Actor, Editor
birthdate = 4.2.1940
birthplace = New York

04.05.2010

```
{{Infobox musical artist
| Name              = Björk
| Img               = Björk Rock en Seine 2007 (2).jpg
| Img_capt          = Björk performance at Rock en Sei
| Img_size          =
| Landscape         =
| Background        = solo_singer
| Birth_name        = Björk Guðmundsdóttir
| Alias             =
| Born              = {{birth date and age|df=yes|196
| Died              =
| Origin            =
| Instrument        = [[Singer|Vocals]], [[Keyboard in           ]], [[harp]]
|Voice_type         = [[Soprano]]<ref>http://www.spiri
| Genre             = [[Electronica]], [[alternative r
| Occupation        = [[Musician]], [[songwriter]], [[record producer|producer]], [[actor|actress]], [[fashion model]]
| Years_active      = 1977–present
| Label             = [[One Little Indian]], [[Elektra Records|Elektra]], [[Atlantic Records Group|Atlantic]], [[Polydor Records|Polydor]], [[Mother Re
[[Smekkleysa]]
| Associated_acts   = [[The Sugarcubes]], [[Thom Yorke]], [[Tappi Tíkarrass]], [[KUKL (group)|KUKL]]
| URL               = [http://www.bjork.com/ www.bjork.com]
}}
```

Björk (Musician)
Occupation = Musician, Actor
Born = 21.12.1965, Reykjavík

```
{{Infobox Prime Minister
|honorific-prefix = <small>
|name             = Gordon
|honorific-suffix = <br /><
|image            = Gordon
|imagesize        = 200px
|alt              = Gordon
|office           = [[Prime
|monarch          = [[Eliza
|term_start       = 27 June
|term_end         =
|predecessor      = [[Tony Blair]]
|successor        =
|office2          = [[Chancellor of the Exchequer]]
|term_start2      = 2 May 1997
|term_end2        = 27 June 2007
|primeminister2   = [[Tony Blair]]
|predecessor2     = [[Kenneth Clarke]]
|successor2       = [[Alistair Darling]]
```

Brown (Prime Minister)
office = Prime Minister of the UK
birth_date = 20.4.1951
birth_place = Govan

```
{{Infobox actor
| image = George Romero, 66
| imagesize = 200px
| caption = Romero in Veni
| birthname = George Andre
| birthdate = {{birth date
| birthplace = [[New York,
| spouse = Christine Forres
| occupation = [[film dire
}}
```

Romero (Actor)
occupation = Actor, Editor
birthdate = 4.2.1940
birthplace = New York

04.05.2010

Richtige Semantik

- Zusammenführen, was zusammen gehört

  - birth_place, placeOfBirth => birthPlace

- Trennen, was nicht zusammen gehört

  - born => birthPlace, birthDate

- Korrektes Klassifikationsschema

  - Wikipedia Kategorien enthalten Zyklen und werden wie Tags verwendet, nicht wie OWL Klassen

Mapping und DBpedia Ontology

- http://mappings.dbpedia.org

    - Termnormalisierung

    - Mapping

    - Ontologieschema

    werden in einem kollaborativen Wikiprozess gepflegt.

Live Extraktion

- Daten werden live mit Wikipedia synchronisiert

- über 100.000 updates pro Tag

- http://dbpedia-live.openlinksw.com/sparql

- Erlaubt das indirekte Editieren von RDF Daten durch Wikipedia

- Gute Datenqualität in Kombination mit kollaborativer Mapping- und Ontologieerstellung

    => hochaktuelle Daten

    => richtige Semantik

Verwendung

- Download auf http://dbpedia.org

- Verfügbar über einen öffentlichen Webservice (OpenLink):

  - SPARQL http://dbpedia.org/sparql

  - Live http://dbpedia-live.openlinksw.com/sparql

  - Linked Data

- Lizensiert als CC-BY-SA (wie Wikipedia)

Vorteile

- Qualitativ hochwertiger Datensatz (durch das Mapping)

- Informationen sind hochaktuell

- leichte programmatische Wiederverwendung

- einfaches erstellen von Mash-Ups durch den Web Service und LinkedData

- Freie Lizenz – bereits hunderte von Applikationen

  - BBC

  - New York Times

  - http://wiki.dbpedia.org/Applications

Vorteile

- Daten sind Ready-to-Use

- Beispiele:

  - Städte-Katalog

  - Daten über führende Politiker der Welt

  - Concept Tagging