Linked Data Profiling
Introduction

22.4.2009
Felix Naumann
Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff
Information systems team

project **ViQTOR**

Katrin Heinrich
DQ Annotation & Assessment

Prof. Felix Naumann

Jens Bleiholder
Data Fusion

Christoph Böhm
DQ Annotation & Assessment

project **fusem**

Paul Führing
Data Profiling & Cleaning

Armin Roth
Information Integration

Peer Data Management Systems

Matching

Information Quality

Service-Oriented Systems

Data Integration for Life Science Data Sources

Ontologies, Profiling

Data Profiling for Schema Management

Mohammed AbuJarour

Frank Kaufer

Jana Bauckmann

project **HumMe**

Jana Bauckmann

Frank Kaufer

Mohammed AbuJarour

Data Profiling & Cleaning

ETL Management

Alexander Albrecht

Felix Naumann | Linked Data Profiling | SoSe 2009
Other courses in this semester

Lectures
- DBS I
- Search engines

Seminars
- Bachelor: Beauty is our Business
- Bachelor: Map/Reduce Algorithms on Hadoop
- Master: Linked Data Profiling
- Forschungsseminar

Bachelorproject
- ETL Management
Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff
Linked data – four principles

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 information.
4. Include links to other URIs, so that they can discover more things.
   □ Many common things are represented in multiple data sets
Linked Data Characteristics

- Linked Data comes as triples
  - Subject, predicate, object
  - URI, property, value
- Linked data is often user generated
- Linked data is free

- Linked data is heterogeneous
- Linked data is dirty
Suggested Reading

- Linked Data by Tim Berners-Lee
  - [http://www.w3.org/DesignIssues/LinkedData.html](http://www.w3.org/DesignIssues/LinkedData.html)

- Interlinking Open Data on the Web by Chris Bizer, Tom Heath, Danny Ayers, and Yves Raimond
  - ESWC 2007
  - [http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkingOpenData.pdf](http://sites.wiwiss.fu-berlin.de/suhl/bizer/pub/LinkingOpenData.pdf)

- And many more at
  - [http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/](http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/)
  - [http://linkeddata.org/](http://linkeddata.org/)
TED talk by Berners-Lee

  - 15 min
DBPedia - Extraction

```json
{{Infobox_Town AT |
    name = Innsbruck |
    image_coa = InnsbruckWappen.png |
    image_map = Karte-tirol-i.png |
    state = {{Tyrol}} |
    regbez = {{Statutory city}} |
    population = 117,942 |
    population_as_of = 2006 |
    pop_dens = 1,119 |
    area = 104.91 |
    elevation = 574 |
    lat_deg = 47 |
    lat_min = 16 |
    lat_edm = N |
    lon_deg = 11 |
    lon_min = 23 |
    lon_edm = E |
    postal_code = 6010-6080 |
    area_code = 0512 |
    licence = I |
    mayor = Hilde Zach |
    website = [http://innsbruck.at] |
}}
```

Country: Austria
State: Tyrol
Administrative region: Statutory city
Population: 117,942
Area: 104.91 km²
Population density: 1,119 /km²
Elevation: 574 m
Coordinates: 47°16' N 11°23' E
Postal code: 6010-6080
Area code: 0512
Licence plate code: I
Mayor: Hilde Zach
Website: www.innsbruck.at
1. Core Datasets

- 274 million triples
- From English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese, Chinese, Russian, Finnish and Norwegian versions of Wikipedia

- 2.6 million things
  - 213,000 persons
  - 328,000 places
  - 57,000 music albums
  - 36,000 films
  - 20,000 companies

How do we know this?

http://wiki.dbpedia.org/Datasets
Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles</td>
<td>DBPResource “rdfsLabel” WikiPTitle</td>
</tr>
<tr>
<td>ShortAbstracts</td>
<td>DBPResource “rdfsComment” Abstract(Text)</td>
</tr>
<tr>
<td>ExtendedAbstracts</td>
<td>DBPResource “dbpsAbstract” Abstract(Text)</td>
</tr>
<tr>
<td>Images</td>
<td>DBPResource “DCrights”, “FOAFdepiction”, “FOAFimg” Abstract(Text)</td>
</tr>
<tr>
<td>LinksToWikiP</td>
<td>DBPResource “FOAFpage” WikiP page</td>
</tr>
<tr>
<td>ArticlesCategories</td>
<td>DBPResource “skosSubject” Category</td>
</tr>
<tr>
<td>ExternalLinks</td>
<td>DBPResource “reference” URL</td>
</tr>
<tr>
<td>InfoBoxes</td>
<td>DBPResource “name”,... (39343 different) Value</td>
</tr>
<tr>
<td>Properties</td>
<td>DBPUrl “rdfsType”; “rdfsLabel” rdfsProperty; label</td>
</tr>
<tr>
<td>Homepages</td>
<td>DBPResource “FOAFHomepage” URL</td>
</tr>
<tr>
<td>GeographicCoordinates</td>
<td>DBPResource 8 different from 3 namespaces Number;URL</td>
</tr>
<tr>
<td>Pagelinks</td>
<td>DBPResource “wikilink” DBPResource</td>
</tr>
<tr>
<td>Persondata</td>
<td>DBPResource “givenname”, “name”, “surname”, “type” plus 3 others very rare  value</td>
</tr>
<tr>
<td>Redirects</td>
<td>DBPResource (almost key, long tail) “redirect” DBPResource</td>
</tr>
<tr>
<td>DisambiguationLinks</td>
<td>DBPResource “disambiguates” DBPResource (almost key, long tail)</td>
</tr>
<tr>
<td>WordnetClasses</td>
<td>DBPResource “wordnettype” WordnetURL</td>
</tr>
<tr>
<td>CategoriesLabels</td>
<td>“Category:”+DBPResource “rdfsLabel” Category label (name, also unique)</td>
</tr>
<tr>
<td>CategoriesSkos</td>
<td>“Category:”+DBPResource “skos-broader”, “skos-prelabel”; “rdfsType” “Category:”+DBPResource; Categorylabe</td>
</tr>
</tbody>
</table>
SELECT DISTINCT P1.PERSONID AS PERSONID, P2.VALUE AS NAME, P3.VALUE AS GIVENNAME, P4.VALUE AS SURNAME
FROM DBPEDIA.PERSONDATA AS P1, DBPEDIA.PERSONDATA AS P2, DBPEDIA.PERSONDATA AS P3, DBPEDIA.PERSONDATA AS P4
WHERE P1.PERSONID = P2.PERSONID
AND P2.PERSONID = P3.PERSONID
AND P3.PERSONID = P4.PERSONID
AND P2.ATTRIBUTE = 'http://xmlns.com/foaf/0.1/name'
AND P3.ATTRIBUTE = 'http://xmlns.com/foaf/0.1/givenname'
AND P4.ATTRIBUTE = 'http://xmlns.com/foaf/0.1/surname');
Some Queries – Profiling Persons

```
SELECT *
FROM DBPEDIA.PIVOTPERSONDATA
WHERE NAME <> GIVENNAME || ' ' || SURNAME;

SELECT COUNT(*)
FROM DBPEDIA.PIVOTPERSONDATA
WHERE NAME = '';

SELECT *
FROM DBPEDIA.PIVOTPERSONDATA AS P1, DBPEDIA.PIVOTPERSONDATA AS P2
WHERE ARESTRINGSSIMILAR(P1.SURNAME, P2.SURNAME, 0.1) = 1
AND ARESTRINGSSIMILAR(P1.GIVENNAME, P2.GIVENNAME, 0.1) = 1
AND P1.PERSONID <> P2.PERSONID
ORDER BY P1.surname, P1.GIVENNAME
```
Companies

- Definition of companies?
  - SELECT DISTINCT TOPIC FROM DBPEDIA.INFOBOXES WHERE ATTRIBUTE = 'companyName'

- Schema?
  - SELECT ATTRIBUTE, COUNT(*) AS SUM
  - FROM DBPEDIA.INFOBOXES
  - WHERE TOPIC IN
  - (SELECT DISTINCT TOPIC FROM DBPEDIA.INFOBOXES WHERE ATTRIBUTE = 'companyName')
  - GROUP BY ATTRIBUTE
  - ORDER BY SUM DESC;
Company attribute distribution
Some more about companies

**Def. 1**: Topics having a `companyName`
- 14292 companies

**Def. 2**: Topics having a `category` that starts with `'compan%'`
- 21753

**Def. 3**: Topics having a `wikiPageUsesTemplate` with value `Template:infobox_company`
- 15491
The queries...

- \[
\text{SELECT COUNT(*) FROM (}
\begin{align*}
&\text{(SELECT DISTINCT I.TOPIC} \\
&\text{FROM DBPEDIA.INFOBOXES I} \\
&\text{WHERE I.ATTRIBUTE = 'companyName'}) \\
&\diamond \text{ INTERSECT} \\
&\text{(SELECT DISTINCT I.TOPIC} \\
&\text{FROM DBPEDIA.INFOBOXES I, DBPEDIA.ARTICLESCATEGORIES C} \\
&\text{WHERE I.TOPIC = C.SUBJECT} \\
&\text{AND UPPER(C.OBJECT) LIKE '%COMPAN%'}) \\
&\diamond \text{ INTERSECT} \\
&\text{(SELECT DISTINCT I.TOPIC} \\
&\text{FROM DBPEDIA.INFOBOXES I} \\
&\text{WHERE I.VALUE = 'Template:infobox_company'})
\end{align*}
\)
1083 different attributes
- H-index 56
- 499 appear only once

Of the 1083 attr., 39 distinct ones contain ‘name’ as substring

273 companies without any name attribute
## Zooming into IBM

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>companyName</td>
<td>International Business Machines Corporation</td>
</tr>
<tr>
<td>companyLogo</td>
<td><img src="http://upload.wikimedia.org/wikipedia/commons/5/51/IBM_logo.svg" alt="IBM Logo" /></td>
</tr>
<tr>
<td>companyType</td>
<td>Public_company</td>
</tr>
<tr>
<td>foundation</td>
<td>Endicott%2C_New_York</td>
</tr>
<tr>
<td>foundation</td>
<td>New_York</td>
</tr>
<tr>
<td>foundation</td>
<td>United_States</td>
</tr>
<tr>
<td>locationCity</td>
<td>Armonk%2C_New_York</td>
</tr>
<tr>
<td>locationCity</td>
<td>New_York</td>
</tr>
<tr>
<td>locationCountry</td>
<td>USA</td>
</tr>
<tr>
<td>slogan</td>
<td>On Demand Business, in demand people</td>
</tr>
<tr>
<td>keyPeople</td>
<td>Samuel_J._Palmisano</td>
</tr>
<tr>
<td>keyPeople</td>
<td>Mark_Loughridge</td>
</tr>
<tr>
<td>keyPeople</td>
<td>Dan_Fortin</td>
</tr>
<tr>
<td>keyPeople</td>
<td>Nick_Donofrio</td>
</tr>
<tr>
<td>industry</td>
<td>Computer_hardware</td>
</tr>
<tr>
<td>industry</td>
<td>Computer_software</td>
</tr>
<tr>
<td>industry</td>
<td>Consultant</td>
</tr>
<tr>
<td>industry</td>
<td>IT_Service_Management</td>
</tr>
<tr>
<td>products</td>
<td>List_of.ibm_products</td>
</tr>
<tr>
<td>revenue</td>
<td>988000000000</td>
</tr>
<tr>
<td>netIncome</td>
<td>10800000000</td>
</tr>
<tr>
<td>numEmployees</td>
<td>386558</td>
</tr>
<tr>
<td>subsid</td>
<td>ADSTAR</td>
</tr>
<tr>
<td>subsid</td>
<td>FileNet</td>
</tr>
<tr>
<td>subsid</td>
<td>Informix</td>
</tr>
<tr>
<td>subsid</td>
<td>Iris_Associates</td>
</tr>
<tr>
<td>subsid</td>
<td>Lotus_Software</td>
</tr>
<tr>
<td>subsid</td>
<td>Rational_Software</td>
</tr>
<tr>
<td>subsid</td>
<td>Sequent_Computer_System%2C_Inc.</td>
</tr>
<tr>
<td>subsid</td>
<td>Tivoli_Systems%2C_Inc.</td>
</tr>
<tr>
<td>homepage</td>
<td><a href="http://www.ibm.com/">IBM Corp.</a></td>
</tr>
<tr>
<td>wikiPageUsesTemplate</td>
<td>Template:infobox_company</td>
</tr>
<tr>
<td>name</td>
<td>secCik</td>
</tr>
<tr>
<td>symbol</td>
<td>hoovers</td>
</tr>
<tr>
<td>secCik</td>
<td>51143</td>
</tr>
<tr>
<td>hoovers</td>
<td>10796</td>
</tr>
<tr>
<td>Template:finance_links</td>
<td></td>
</tr>
</tbody>
</table>
Schemata of IBM, Microsoft, and Oracle

IBM
- companyLogo
- companyName
- companyType
- foundation
- homepage
- hoovers
- industry
- keyPeople
- locationCity
- locationCountry
- name
- netIncome
- numEmployees
- othersUse6Property
- products
- revenue
- secCik
- slogan
- subsid
- symbol
- wikiPageUsesTemplate

Microsoft
- _percent_22Property
- areaServed
- assets
- companyLogo
- companyName
- companyType
- equity
- foundation
- founder
- homepage
- industry
- keyPeople
- location
- marketCap
- netIncome
- numEmployees
- operatingIncome
- products
- relatedInstance
- revenue
- slogan
- wikiPageUsesTemplate

Oracle_Corporation
- companyLogo
- companyName
- companyType
- founder
- homepage
- industry
- keyPeople
- locationCity
- name
- numEmployees
- products
- revenue
- slogan
- wikiPageUsesTemplate
The problem – a schema and data mess

- Triples and ill-defined templates invite disaster.
- Extract attribute values from various text fields in the source
  - 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
  - Sloppy data entry: Data value are neither standardized nor normalized
    - Revenue attribute in DBPedia may contain different units, different currencies, and different number-formats.
      - 1.64 billion USD vs. $1640 m vs. 1,6 vs. more than one million Euro in 2006
      - And lots of other stuff: ? or Undisclosed or Image:green_up.png or Assets exceed £4 billion GBP or Wal-Mart or € bn (as of 2004) or http://www.credit-suisse.com/investors/en/reports/2007_results_q4.jsp
LOD Gear: System for building and maintaining high quality Linked-Open-Data

Query: Name: “American Express”

Name: American Express Co
CEO: Kenneth I. Chenault
Address: 200 Vesey Street, 50th floor, New York, NY, 10285

DBPedia
Freebase
SEC
etc…

Integration Engine
Explore/Extract, Scrub, Resolve, Map, Fuse

DaaS Application

Financial Companies Search Interface

Financial Companies Index

Variant Generation

Financial Companies

Felix Naumann | Linked Data Profiling | SoSe 2009
Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff
Three themes

- Data heterogeneity
- Semantic heterogeneity
- Object heterogeneity
Project Goals: Profiling Data Heterogeneity

- In DBPedia there are companies, people, artists, buildings, and many more different concepts.

- Goal:
  - find out such concepts or cluster of objects
  - display them with useful / cool statistics
  - allow inter- concept / cluster comparisons
  - allow comparison with existing classifications
Given existing concepts / classes of objects, e.g. all subjects that use a specific Wikipedia template.

Goal:
- Find out interesting / cool statistics
- Find Patterns
  - longitude=73° → time zone=UTC-5
  - company name attribute → industry attribute
- Explore Schema Matchings
  - suggestions
  - what if name and family name were merged
- Explore Join Paths
  - location – river / city / country
Project Goals: Profiling Object Heterogeneity

- Given existing concepts / classes of objects, e.g. all subjects that use a specific Wikipedia template.

- Goal:
  - How much information do objects provide?
  - How do attribute values correspond to Wikipedia Texts?
  - What can we find out by looking at external resources?
  - Rank objects ...
Infrastructure

- mut/sokar.hpi.uni-potsdam.de
- DB2 database **lodprof** (requires tuning!)
  - DBPEDIA.INFOBOXES
  - DBPEDIA.ARTICLES_CATEGORIES
  - DBPEDIA.CATEGORIES_LABEL
  - DBPEDIA.DISAMBIGUATIONLINKS
  - DBPEDIA.LINKSTOWIKIPEDIA
  - DBPEDIA.ONTOLOGY_INFOBOXES
  - DBPEDIA.ONTOLOGY_TYPES
  - DBPEDIA.PAGELINKS
  - DBPEDIA.PERSONDATA
  - DBPEDIA.SHORTABSTRACTS
  - DBPEDIA.TITLES

- User/Pwd: see whiteboard
- Java 1.6
- Google Web Toolkit
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>%22Dm%22_Is_for_Burglar</td>
<td>edit</td>
<td>11281665</td>
</tr>
<tr>
<td>%22Dm%22_Is_for_Burglar</td>
<td>preacedBy</td>
<td>%22Mr%22_Is_for_Apel</td>
</tr>
<tr>
<td>%22Dm%22_Is_for_Burglar</td>
<td>followers</td>
<td>%22Mr%22_Is_for_Corge</td>
</tr>
<tr>
<td>%22Dm%22_Is_for_Burglar</td>
<td>wikipediaTemplate</td>
<td>Template:inbook_book</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>characterName</td>
<td>Ran</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>nameplate</td>
<td>Roh_Tan</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>years</td>
<td>1985, 1984, 1987</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>occupation</td>
<td>Farmer</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>home</td>
<td>Spain</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>daughters</td>
<td>Tina</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>wikipediaTemplate</td>
<td>Template:inbook_readers_character_2</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>name</td>
<td>回顾 The Movie</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>type</td>
<td>DVD</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>url</td>
<td>%22Blg%22_Ran</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>cover</td>
<td>BTsT.max.jpg</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>release</td>
<td>January 1, 1986 - DVD</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>genre</td>
<td>Hard摇滚</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>length</td>
<td>1:11:09</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>label</td>
<td>Columbia Records</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>producer</td>
<td>Steve Golden</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>sub2End</td>
<td>Live Ripper (1)</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>nextMovie</td>
<td>The True Meaning of %22Brotherhood%22337</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>wikipediaTemplate</td>
<td>Template:inbook_movie_dvd</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>milestoneProperty</td>
<td>Pleasure US - Visit to Kanama -</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>milestoneProperty</td>
<td>Pleasure US - Ctrl &amp; User -</td>
</tr>
<tr>
<td>%22Blg%22_Ran</td>
<td>wikipediaTemplate</td>
<td>Template:inbook_rhonege</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>name</td>
<td>By My Side</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>cover</td>
<td><a href="http://upload.wikimedia.org/wikipedia/commons/4/42/By_Your_Side_Cover.jpg">http://upload.wikimedia.org/wikipedia/commons/4/42/By_Your_Side_Cover.jpg</a></td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>artist</td>
<td>Yoko</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>artist</td>
<td>Yoko</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>release</td>
<td>2002-30</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>format</td>
<td>CD single</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>format</td>
<td>Digital Remix</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>genre</td>
<td>Gangsta Tap</td>
</tr>
<tr>
<td>%22Blg%22_My_Side%22</td>
<td>length</td>
<td>3:39</td>
</tr>
</tbody>
</table>
Overview

- Introduction to team
- Linked Data
- Linked Data Exploration
  - My IBM Experience
- Linked Data Profiling
  - The Project Goals
- Organizational Stuff
Homework!

- Explore DBPedia (and Wikipedia)!
  - Appreciate problem and profiling need
  - Next week: Mr. DBPedia is here
    - 29.4. 13 – 17 Uhr

- See TED-talk by Tim Berners-Lee

- Choose problem (email with top 2 choices)
  - Infrastructure (6x 1/6 person)
  - Domain heterogeneity (2 persons)
  - Schematic heterogeneity (2 persons)
  - Object heterogeneity (Discovery per Object) (2 persons)

- Identify solution directions
Questions, wishes, ... 

- Now, or ...
- Office: A.1-13
- Consultations: Tuesdays 15-16 Uhr
  or by arrangement
- Email: naumann@hpi.uni-potsdam.de
- Phone: (0331) 5509 280

The end.