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](http://www.w3.org/DesignIssues/LinkedData.html), 2006
The RDF Data Model

```
ns:anja rdf:type foaf:Person
    foaf:name Anja Jentzsch
    foaf:based_near dbpedia:Berlin
```
Identifying Data Items using URIs

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

dbpedia:Berlin = http://dbpedia.org/resource/Berlin
Dereferencing URIs over the Web

```
ns:anja rdf:type foaf:Person
  foaf:name Anja Jentzsch
  foaf:based_near dbpedia:Berlin
    dp:population 3.499.879
    skos:subject dp:Cities_in_Germany
```
Dereferencing URIs over the Web

- ns:anja
  - foaf:name: Anja Jentzsch
  - foaf:based_near: dbpedia:Berlin
- dbpedia:Hamburg
- dbpedia:Muenchen
- dp:Cities_in_Germany
- dbpedia:Muenchen
- dbpedia:Hamburg
- dp:population: 3,499,879
RDF Representation Formats

- **RDF/XML**
  
  ```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:RDF>
  ```

- **RDF N-Triples**
  
  ```turtle
  @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
  @prefix foaf: <http://xmlns.com/foaf/0.1/> .
  foaf:name "Anja Jentzsch" .
  ```

- **RDF N3**
RDF Representation Formats

- `<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 July 2009)
- 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

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

![Graph showing growth in numbers over years](visualisation.png)
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/
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
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 is a joint project with the following goals:

- Extracting structured information from Wikipedia
- Publishing 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

Berlin (English pronunciation /ˈbɛrlɪn/; German pronunciation: [ˈbeːrlɪn] (listen)) 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.

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). Berlin in the 1920s was the third largest municipality in the world. 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). Following German reunification in 1990, the city regained its status as the capital of Germany, hosting 147 foreign embassies.

Berlin is a world city of culture, politics, media, and science. 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 and is a popular tourist destination. 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. Its urban settings and historical legacy have made it a popular location for international film productions. The city is well known for its festivals, diverse architecture, nightlife, contemporary arts, public transportation networks and a high quality of living.
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 .


- 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

- displays Wikipedia data on map
- aggregates different data sources
Faceted browsing and free text search

Faceted Wikipedia Search

- Item type
  - Skyscraper (3)
  - Place (3)
  - Building (3)

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

- Architect
  - Adrian Smith (architect) (1)

- Floor area (m²)
  - 273349 (1)
  - 278707 (1)

- Building started in year

Jin Mao Tower
The Jin Mao Tower is an 88-story landmark super-tall 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 133 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 (spr 和記實業有限公司), which constructed the building. Hopewell Holdings Limited and is headquartered in the building and its CEO, Sir Gordon Wu Ying Shuang, has his office on the top floor of the building.
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.

http://spotlight.dbpedia.org
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

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)

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 excerps:
Companies in DBpedia

- Def. 1: Subject having a predicate `companyName`
  - \( \rightarrow 14,292 \)

- Def. 2: Subject having a category that starts with 'compane'
  - \( \rightarrow 21,753 \)

- Def. 3: Subject having a `wikiPageUsesTemplate` with value `Template:infobox_company`
  - \( \rightarrow 15,491 \)
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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>20617</td>
<td>13355</td>
</tr>
<tr>
<td>products</td>
<td>18176</td>
<td></td>
</tr>
<tr>
<td>pageTitleTemplate</td>
<td>18048</td>
<td>2036</td>
</tr>
<tr>
<td>keyPeople</td>
<td>17836</td>
<td></td>
</tr>
<tr>
<td>industry</td>
<td>16822</td>
<td></td>
</tr>
<tr>
<td>foundation</td>
<td>15826</td>
<td></td>
</tr>
<tr>
<td>homepage</td>
<td>14476</td>
<td></td>
</tr>
<tr>
<td>companyType</td>
<td>13433</td>
<td></td>
</tr>
<tr>
<td>companyLogo</td>
<td>13355</td>
<td></td>
</tr>
<tr>
<td>numEmployees</td>
<td>12506</td>
<td></td>
</tr>
<tr>
<td>revenue</td>
<td>11751</td>
<td></td>
</tr>
<tr>
<td>locationCity</td>
<td>9006</td>
<td></td>
</tr>
<tr>
<td>locationCountry</td>
<td>6207</td>
<td></td>
</tr>
<tr>
<td>companySlogan</td>
<td>5030</td>
<td></td>
</tr>
<tr>
<td>areaSlogan</td>
<td>4098</td>
<td></td>
</tr>
<tr>
<td>relatedInstance</td>
<td>3212</td>
<td></td>
</tr>
<tr>
<td>type</td>
<td>2615</td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>2557</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>2284</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>2152</td>
<td></td>
</tr>
<tr>
<td>founder</td>
<td>2054</td>
<td></td>
</tr>
<tr>
<td>subsid</td>
<td>2036</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>1663</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>1597</td>
<td></td>
</tr>
<tr>
<td>founder</td>
<td>1232</td>
<td></td>
</tr>
<tr>
<td>subsid</td>
<td>1141</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>1087</td>
<td></td>
</tr>
<tr>
<td>founder</td>
<td>960</td>
<td></td>
</tr>
<tr>
<td>subsid</td>
<td>925</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>904</td>
<td></td>
</tr>
<tr>
<td>founder</td>
<td>896</td>
<td></td>
</tr>
<tr>
<td>subsid</td>
<td>885</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>808</td>
<td></td>
</tr>
<tr>
<td>founder</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>subsid</td>
<td>503</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>foundation</td>
<td>429</td>
<td></td>
</tr>
<tr>
<td>foundation</td>
<td>422</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>419</td>
<td></td>
</tr>
</tbody>
</table>

Profiling Linked Open Data | HPI 2013 | Anja Jentzsch
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-lingual 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
Challenges:

- Heterogeneity
  - Loose structure: Things have different predicate sets
  - 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

Application Layer

Data Access, Integration and Storage Layer

Web of Linked Data

Publication Layer
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
- ...

Profiling Linked Open Data | HPI 2013 | Anja Jentzsch
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 sets
  Average 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)
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
Suite of methods ranging from:
- Domain level (clustering, labeling)
- Schema level (matching, disambiguation)
- Data level (data type detection, pattern detection, value distribution)
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
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)
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
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
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
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
- Apriori Algorithm, Agrawal and Srikant, 1994
  - media cluster example:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Confidence</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>genre, isbn ⇒ author</td>
<td>0.99</td>
<td>0.67</td>
</tr>
<tr>
<td>isbn ⇒ author</td>
<td>0.92</td>
<td>0.66</td>
</tr>
<tr>
<td>isbn ⇒ author, genre</td>
<td>0.83</td>
<td>0.66</td>
</tr>
<tr>
<td>author, genre ⇒ isbn</td>
<td>0.70</td>
<td>0.66</td>
</tr>
<tr>
<td>author ⇒ isbn</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>author ⇒ genre, isbn</td>
<td>0.58</td>
<td>0.67</td>
</tr>
</tbody>
</table>

- Conclusion:
  - genre, isbn, author together form part of an entity’s schema
    - Assumption: complement each other
  - distinct semantics
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
Subject X holds link to Subject Y via predicate $X \xrightarrow{A} Y$

$X \xrightarrow{A} Y$ and $Y \xleftarrow{B} X$ , then A and B are inverse links.

Example:

<table>
<thead>
<tr>
<th>PredicateA</th>
<th>PredicateB</th>
<th>Corr Coef</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>before</td>
<td>after</td>
<td>0.239</td>
<td>28856</td>
</tr>
<tr>
<td>sisterStations</td>
<td>sisterStations</td>
<td>0.749</td>
<td>7494</td>
</tr>
<tr>
<td>precededBy</td>
<td>followedBy</td>
<td>0.830</td>
<td>7097</td>
</tr>
<tr>
<td>spouse</td>
<td>spouse</td>
<td>0.322</td>
<td>1964</td>
</tr>
<tr>
<td>before</td>
<td>before</td>
<td>-0.003</td>
<td>738</td>
</tr>
<tr>
<td>star</td>
<td>exoplanet</td>
<td>0.895</td>
<td>188</td>
</tr>
</tbody>
</table>

Conclusion:

- Redundancy of e.g. before/after and sisterStations
  - Fuse with ProLOD
- Misuse of before
  - Exclude before with ProLOD
- (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 $\rightarrow$ determine (normalized) patterns
  - Integers, Decimals $\rightarrow$ 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 ...)

Data + Views

Profiling

continuous process

Understanding

Metadata
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

ontology:city

rdfs:subClassOf

ontology:capital

example.com/Munich

example.com/Berlin

example.com/Alice

example.com/Bob

foo.org/place#Lyon

foo.org/place#Berlin

foo.org/person#Alice

foo.org/person#Bob

foo.org/place#Paris

bar.net/entity:Alice

bar.net/entity:Bob

owl:sameAs

rdf:type

worksIn

livesIn
Connected Datasets for voiD?

example.com/Alice
example.com/Bob

example.com/Munich
example.com/Berlin

foo.org/place#Lyon
foo.org/place#Berlin
foo.org/place#Paris

bar.net/entity:Alice
bar.net/entity:Bob

ontology:city
ontology:capital
rdfs:subClassOf

rdf:type
worksIn
livesIn
owl:sameAs
Conceptual Datasets for voiD?

city dataset
capital dataset

Profiling Linked Open Data | HPI 2013 | Anja Jentzsch
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
Keyword-based search for relevant instances:

- Randomly select subset of individuals 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
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

- 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


