UNIVERSITÄT MANNHEIM

Learning Conflict Resolution Strategies for Cross-Language Wikipedia Data Fusion

Volha Bryl, Christian Bizer

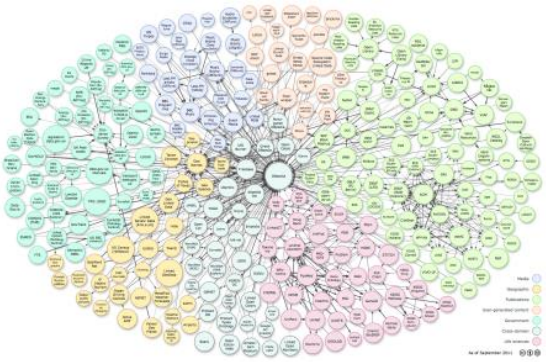
Data and Web Science Research Group
University of Mannheim
Germany

WebQuality @ WWW'2014, Seoul, Korea, April 7, 2014

Outline

1. Motivation
   - Linked Open Data fusion
   - Wikipedia/DBpedia data fusion
2. Extracting provenance metadata
3. Data fusion with Sieve
4. Learning data fusion policies
5. Cross-language DBpedia use case
6. Conclusion

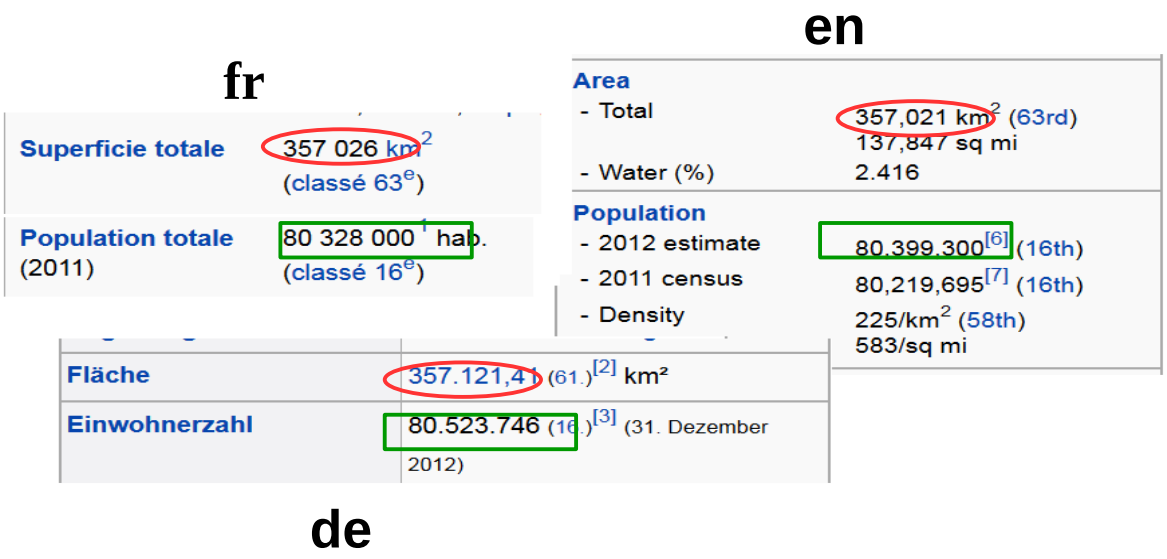
Motivation: Linked Open Data Integration



- **LOD** – publishing and interlinking open datasets on the web
  - Tens of billions of facts (*RDF* subject-predicate-object *triples*)
- Huge potential for applications
- Problem: **varying quality, lack of data consistency**
- Solution: **data integration**
- Our focus: **data fusion**
  - *Create a consistent representation of a real-world entity based on multiple heterogeneous data sources*
  - Challenge: **conflict resolution** for improving data quality

Motivation: Fusing Wikipedia Data

- Wikipedia contains **lots of data conflicts** across languages
- **Improving** its **quality** is crucial
- Identity resolution is solved by inter-language links



- Our focus is **DBpedia**: Wikipedia's structured twin

DBpedia: Wikipedia's Structured Twin

- Extracts structured, multilingual, cross-domain knowledge from Wikipedia
- Crowd-sourced community project: http://dbpedia.org

- Provides **querying and search capabilities over Wikipedia data**
- Follows Linked Open Data principles
- Data is freely available, software is open-source

- Started in 2006, currently at version 3.9 (September 2013)
- 119 languages, **2.46 billion triples**, 12.6 million unique things
- LOD cloud's interlinking hub

DBpedia: Data Conflicts

- Querying DBpedia: What is the population of Seoul?

Seoul	dbp-ont:populationTotal	"10,447,719"@bg "9,794,304"@ca "10,400,000"@cs "10,464,051"@el "10,581,728"@en "9,794,304"@eu "10,464,051"@id "14,794,304"@it "10,528,774"@ko "10,581,728"@pt "10,464,051"@ru "10,581,728"@sl "24,500,000"@tr

DBpedia: Data Conflicts

- Querying DBpedia: What is the population of Seoul?

Seoul	dbp-ont:populationTotal	"10,447,719"@bg "**9,794,304**"@ca "10,400,000"@cs "**10,464,051**"@el "**10,581,728**"@en "**9,794,304**"@eu "**10,464,051**"@id "14,794,304"@it "10,528,774"@ko "**10,581,728**"@pt "**10,464,051**"@ru "**10,581,728**"@sl "24,500,000"@tr

DBpedia: Data Conflicts

- Querying DBpedia: What is the population of Seoul?

Seoul	dbp-ont:populationTotal	"10,447,719"@bg
		"**9,794,304**"@ca
		"10,400,000"@cs
		"**10,464,051**"@el
	Edited 2013-01-22T09:43:20Z	"**10,581,728**"@en
		"**9,794,304**"@eu
		"**10,464,051**"@id
		"14,794,304"@it
		"10,528,774"@ko
		"**10,581,728**"@pt
	Edited 2012-09-28T11:59:15Z	"**10,464,051**"@ru
		"**10,581,728**"@sl
		"24,500,000"@tr

DBpedia: Extracting Provenance Metadata

- **Provenance is crucial for data fusion**
  - Allows assessing data quality, e.g. decide whether the fact is up-to-date or comes from the trusted source or author
- **No provenance metadata** provided for DBpedia at the moment

- Idea
  - **Extract** provenance metadata **from Wikipedia revision history**

- Implementation
  - https://github.com/VolhaBryl/DBpedia-provenance
  - Extraction performed for 610K populated places in 10 languages

DBpedia: Extracting Provenance Metadata

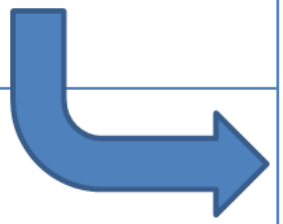
- Two types of metadata extracted: when and how many times, and by whom the fact was edited
- **Change history**: last edit timestamp of a triple, number of edits
  - Retrieved from **Wikipedia revision dumps**
    - April 2013, corresponds to DBpedia 3.9 release
  - Challenge: revision dumps are **huge**
    - e.g. **>6Tb** for English, **>2Tb** for German
  - We extracted metadata for geographical entities for 10 top languages
    - 425K entities for English, ~150K for other languages
- **Author**: reputation-related metadata
  - edit count, registration date, blocked or not, etc.
  - Retrieved via **MediaWiki API**

DBpedia: Extracting Provenance Metadata

Provenance metadata per triple

Wikipedia revision dumps (nl.wiki)

```
<page>
  <title>Mannheim</title>
    <ns>0</ns>
    <id>92489</id>
    <revision>
      <id>39100553</id>
      <parentid>38723407</parentid>
      <timestamp>2013-10-03T18:20:49Z</timestamp>
      <contributor>
        <username>Bean 19</username>
        <id>5048</id>
      </contributor>
      <text> ... </text>
    </revision>
    ...
</page>
```



Example of extracted provenance metadata

```
ru:Mannheim:populationTotal:1
            author      EmausBot
            autheditcnt 1,136,639
            propeditcnt 3
            authregdate 2009-12-18T02:08:09Z
            lastedit    2011-12-22T00:50:21Z
            authbot     true
nl:Mannheim:populationTotal:1
            author      Joopwiki
            autheditcnt 106,899
            propeditcnt 1
            authregdate 2007-04-05T08:54:19Z
            lastedit    2007-12-09T16:41:06Z
            authsysop   true
```

DBpedia: Extracting Provenance Metadata

Edit traces

```
http://de.dbpedia.org/resource/Mannheim          dbpedia-owl:areaTotal    21
            2011-05-20T15:18:04Z    567779       Philipp Wetzlar          144.0;
            2011-05-20T15:17:48Z    84.171.143.73                         ;
            2009-11-10T10:08:05Z    137969       Revvar                   144.0;
            2009-11-10T10:07:52Z    87.160.194.166                        ;
            2009-02-16T18:52:35Z    236879       Kallistratos             144.0;
            2009-02-16T18:48:55Z    217.227.91.34                         3.00450154E8;
            2008-07-12T14:51:57Z    4758         Frank-m                  144.0;
            2008-07-12T13:48:55Z    89.48.60.16                           178.0;
            2008-06-21T14:13:04Z    262856       Wo st 01                 144.0;
            2008-06-21T14:02:28Z    89.48.19.45                           178.0;
            ...
            2006-12-08T18:45:24Z    34480        S.K.                     144.0;

http://pl.dbpedia.org/resource/Mannheim          dbpedia-owl:areaTotal    5
            2009-07-24T10:27:07Z    19873        Malarz pl                1.4496E8;
            2009-07-24T10:20:47Z    31007        MalarzBOT                ;
            2007-01-12T14:52:21Z    27694        FilMys                   1.4496E8;
            2007-01-11T13:34:59Z    27694        FilMys                   ;
            2007-01-11T13:34:08Z    27694        FilMys                   3.1043E8;
```

Data Fusion with Sieve

- Our starting point

  Sieve – Linked Data Quality Assessment and Fusion tool

  http://sieve.wbsg.de/

- Functionality
  - Input: RDF data + provenance metadata
  - User creates an XML specification with
    - **quality assessment metrics** (e.g. recency or trust in source)
    - **conflict resolution functions** (e.g. vote, take maximum, average, most recent, most trusted source)
  - Fused dataset is produced

Data Fusion with Sieve by Example

<dbp:Amsterdam> <dbp-ont:populationTotal> "820654" <**en**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "790044" <**ru**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "762" <**es**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "57" <**es**.wikipedia.org/wiki/Amsterdam:populationTotal:2> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "799406" <**nl**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "1364422" <**pt**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "758198" <**ca**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Amsterdam> <dbp-ont:populationTotal> "820654" <**it**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .
<dbp:Vienna> <dbp-ont:populationTotal> "1731236" <**en**.wikipedia.org/wiki/Vienna:populationTotal:1> .
<dbp:Vienna> <dbp-ont:populationTotal> "1730278" <**ru**.wikipedia.org/wiki/Vienna:populationTotal:1> .
<dbp:Vienna> <dbp-ont:populationTotal> "1731236" <**es**.wikipedia.org/wiki/Vienna:populationTotal:1> .
<dbp:Vienna> <dbp-ont:populationTotal> "1680266" <**ca**.wikipedia.org/wiki/Vienna:populationTotal:1> .
<dbp:Vienna> <dbp-ont:populationTotal> "1731236" <**it**.wikipedia.org/wiki/Vienna:populationTotal:1> .
<dbp:Vienna> <dbp-ont:populationTotal> "1731286" <**fr**.wikipedia.org/wiki/Vienna:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2234105" <**en**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2268265" <**ru**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2257981" <**es**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2257981" <**nl**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2257981" <**nl**.wikipedia.org/wiki/Paris:populationTotal:2> .
<dbp:Paris> <dbp-ont:populationTotal> "2211297" <**pt**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2257981" <**it**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "2243833" <**fr**.wikipedia.org/wiki/Paris:populationTotal:1> .
<dbp:Paris> <dbp-ont:populationTotal> "10413386" <**de**.wikipedia.org/wiki/Paris:populationTotal:1> .

INPUT

23 values,
only 3 needed

Data Fusion with Sieve by Example

```
<QualityAssessment name="Multilingual DBpedia quality assessment scores">
    <AssessmentMetric id="sieve:english">
        <ScoringFunction class="ScoredPrefixList">
            <Param name="list" value="http://en.wikipedia.org"/>
        </ScoringFunction>
    </AssessmentMetric>
    <AssessmentMetric id="sieve:recency">
        <ScoringFunction class="TimeCloseness">
            <Param name="timeSpan" value="500"/>
            <Input path="?GRAPH/dbpedia-meta:lastedit"/>
        </ScoringFunction>
    </AssessmentMetric>
</QualityAssessment>

<Fusion name="Multilingual DBpedia quality assessment scores">
    <Class name="dbpedia-owl:PopulatedPlace">
        <Property name="dbpedia-owl:country">
            <FusionFunction class="KeepFirst" metric="sieve:english"/>
        </Property>
        <Property name="dbpedia-owl:populationTotal">
            <FusionFunction class="KeepFirst" metric="sieve:recency"/>
        </Property>
    </Class>
</Fusion>
```

Sieve specification: keep most recent population value

Data Fusion with Sieve by Example

RESULT (selected most recent population value)

<dbp:Vienna**>** <dbp-ont:populationTotal> "1730278" <**ru**.wikipedia.org/wiki/Vienna:populationTotal:1> .

<dbp:Paris> **<**dbp-ont:populationTotal> "2257981" <**nl**.wikipedia.org/wiki/Paris:populationTotal:2> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "57" <**es**.wikipedia.org/wiki/Amsterdam:populationTotal:2> .

Data Fusion with Sieve by Example

RESULT (selected most recent population value)

<dbp:Vienna> <dbp-ont:populationTotal> "1730278" <**ru**.wikipedia.org/wiki/Vienna:populationTotal:1> .

<dbp:Paris> <dbp-ont:populationTotal> "2257981" <**nl**.wikipedia.org/wiki/Paris:populationTotal:2> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "**57**" <**es**.wikipedia.org/wiki/Amsterdam:populationTotal:2> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "820654" <**en**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "790044" <**ru**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "762" <**es**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "57" <**es**.wikipedia.org/wiki/Amsterdam:populationTotal:2> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "799406" <**nl**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "1364422" <**pt**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "758198" <**ca**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

<dbp:Amsterdam> <dbp-ont:populationTotal> "820654" <**it**.wikipedia.org/wiki/Amsterdam:populationTotal:1> .

Was it wrong to keep the most recent value?..

Learning Conflict Resolution Strategies

- Problem
  - In Sieve fusion functions for each property are **manually defined**
    - ⇒ good understanding of input data required
    - ⇒ optimal result is not guaranteed

- Solution
  - **Fusion Policy Learner**
    - Extension of Sieve for **automatically learning conflict resolution strategies** based on a gold standard
    - http://sieve.wbsg.de/FPL.html

Learning Conflict Resolution Strategies

- **Fusion Policy Learner**
  - Extension of Sieve for **automatically learning conflict resolution strategies** based on a gold standard
  - http://sieve.wbsg.de/FPL.html

- **Learning algorithms**
  - *Numeric properties*
    - Minimize the mean absolute error or
    - Maximize the number of *correct* values
      - Where a *correct* value deviates from the gold standard value by no more than a predefined threshold (e.g. 5%)
  - *Nominal properties* (strings or URIs)
    - Maximize the number of exact matches

Sieve Fusion Policy Learner: Example

Learn which fusion functions to use

```
<Parameters>
   <SelectionMethod name="MinAbsError"/>
</Parameters>
<Input>
    <GoldStandard>gold\3-capitals.gold.nt</GoldStandard>
    <dumpLocation>dumps-3cities</dumpLocation>
    <SieveExec>c:\ldif-0.5.2\bin\ldif.bat</SieveExec>
</Input>
<Output>
    <SieveSpec>sieve-optimal\sieve_optimal.xml</SieveSpec>
    <FPLReport valmatrix = "true">FPL_report.txt</FPLReport>
</Output>
...
    <Fusion name="Multilingual DBpedia quality assessment scores">
        <Class name="dbpedia-owl:PpopulatedPlace">
            <Property name="dbpedia-owl:populationTotal">
                <FusionFunction class="Average"/>
                <FusionFunction class="Maximum"/>
                <FusionFunction class="Voting"/>
                <FusionFunction class="KeepFirst" metric="sieve:recency"/>
                <FusionFunction class="KeepFirst" metric="sieve:english"/>
                <FusionFunction class="KeepFirst" metric="sieve:authactivity"/>
                <FusionFunction class="KeepFirst" metric="sieve:propactivity"/>
                <FusionFunction class="KeepLast" metric="sieve:authage"/>
            </Property>
        </Class>
    </Fusion>
```

Based on the gold standard

Select one of these functions

Input specification

Fusing DBpedia Data

- Top 10 DBpedia language editions, 610,017 entities
  - 30% described in > 3 languages
- Gold standard: GeoNames - www.geonames.org

Property	Dataset	Size	Fusion policy	Error, %	Error, %, en.dbp
populationTotal	cities1000-Germany	7,330	MostFrequent	0.3029	0.6796
populationTotal	cities1000-Netherlands	493	Maximum	2.1933	3.5714
populationTotal	countries	243	Maximum	2.1646	6.3485
populationTotal	cities1000-Brazil	1,247	MostActive property	2.2727	2.6913
country	cities1000-Italy	1,078	MostFrequent	0	1.206
country	cities1000-Brazil	1,119	MostActive author	9.8302	30.9205
country	cities1000-Germany	7,638	MostFrequent	0.0131	0.6415

* selection method for population: minimize mean absolute error

Fusing DBpedia Data

- Top 10 DBpedia language editions, 610,017 entities
  - 30% described in > 3 languages
- Gold standard: GeoNames - www.geonames.org

Property	Dataset	Size	Fusion policy	Error, %	Error, %, en.dbp
populationTotal	cities1000-Germany	7,330	MostFrequent	0.3029	0.6796
populationTotal	cities1000-Netherlands	493	Maximum	2.1933	3.5714
populationTotal	countries	243	Maximum	2.1646	6.3485
populationTotal	cities1000-Brazil	1,247	MostActive property	2.2727	2.6913
country	cities1000-Italy	1,078	MostFrequent	0	1.206
country	cities1000-Brazil	1,119	MostActive author	9.8302	30.9205
country	cities1000-Germany	7,638	MostFrequent	0.0131	0.6415

* selection method for population: minimize mean absolute error

Based on provenance metadata

Summary

- Motivation
  - Data integration is crucial for boosting the quality and usage of LOD
- Objective
  - Fusing Wikipedia/DBpedia data across languages
- Starting point
  - Sieve, LOD quality assessment and fusion tool

- Results
  - Fusion Policy Learner extension of Sieve for automatically learning optimal conflict resolution strategies
  - Fusing data about 610K populated places from 10 DBpedia language editions
  - Framework for DBpedia provenance metadata extraction

Future Work

- Experimenting with other learning techniques
  - regression for numerical values
  - decision trees to learn complex fusion strategies,
    e.g. choose the most recent among the most frequent values
  - active learning when no or not enough labeled data available
- DBpedia use case
  - Is there a cross-domain up-to-date gold standard?
  - Gap filling, conflict resolution and data debugging on a large scale
- Other LOD use cases
  - Allows DBpedia to be used for training