Dr. Crowdsourse
or: How I Learned to Stop Worrying and Love Web Data

Invited Talk @ BEWEB 2011

25.3.2011
Felix Naumann
Dr. Strangelove – for the small frys...
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
A brief history of data
Linked Data & Data Spaces – a database guy’s PoV

- Some Schema
- Integrated
- Ad-hoc
- Data quality
- High Accessibility

Dataspaces / Data integration

Relational databases

Linked data

Semantic 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 information**.

4. Include **links to other URIs**, so that they can discover more things.
   - Many common things are represented in multiple data sets!

**The Good**
- Comes as triples
  - `S: http://.../Uppsala`
  - `P: location`
  - `O: http://.../Sweden`
- Often user generated
- Nice domains
- Free

**The Bad**
- Voluminous
- Heterogeneous

**The Ugly**
- Dirty, inconsistent, sparse
Linked Data Graph

Linked Data Graph

DBpedia – Extraction

{(Infobox Non-profit
 | Non-profit_name  = IEEE
 | Non-profit_logo  = [[Image:IEEE logo.svg|200px]]
 | Non-profit_type  = Professional Organization
 | founded_date     = January 1, 1963
 | founder          =
 | location         =
 | origins          = Merger of the American Institute of Electrical Engineers and
 | key_people       = Mr. Pedro A. Ray, Current President
 | area_served      = Worldwide
 | focus            = Electrical, Electronics, and Information Technology [http://w
 | /visionmission.html]
 | method           = Industry standards, Conferences, Publications
 | revenue          = US$330 million
 | endowment        =
 | num_volunteers   =
 | num_employees    =
 | num_members      = 395,000+
 | owner            =
 | Non-profit_slogan =
 | tax_exempt       =
 | dissolved        =
 | footnotes        =
)}
### DBpedia statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>en</th>
<th>de</th>
<th>fr</th>
<th>es</th>
<th>de</th>
<th>pl</th>
<th>fi</th>
<th>sv</th>
<th>it</th>
<th>zh</th>
<th>nl</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>Titles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short Abstracts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended Abstracts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links to Wikipedia Article</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Articles Categories</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>External Links</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infobuses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Properties</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBpedia Ontology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontology Infobuses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ontology Types</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homepages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geographic Coordinates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pagelinks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persondata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Redirects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disambiguation Links</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **672 million triples**
- **286 million English**
- From 97 languages of Wikipedia
- **3.5 million things**
- 364,000 persons
- 462,000 places
- 99,000 music albums
- 54,000 films
- 16,500 video games

http://wiki.dbpedia.org/Datasets

Felix Naumann | Cleansing Web Data | BEWEB 2011
And more sources

- Government data
  - www.data.gov
  - data.gov.uk
  - ec.europa.eu/eurostat
- Finance / business data
- Scientific databases
  - www.uniprot.org
  - skyserver.sdss.org
- The Web
  - HTML tables and lists
  - General sources: DBpedia, freebase, ...
  - Domain-specific sources: IMDB, Gracenote, isbndb, ...
- ...

“Raw data now!”
Use cases

- General purpose integration: Create rich knowledge bases
  - Semantic Web
  - Improved search / question answering
  - Link creation and data enrichment
  - Cleansing: data correction and validation

- Domain specific integration
  - Creation of high quality data sets: Complete & accurate
  - Enhancement of organization-internal data
  - Create reference data sets
  - Mashups

KILLER APP?
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
The novel has become famous for its portrayal of pervasive government surveillance and control, and government's increasing encroachment on the rights of the individual. Since its publication, many of its terms and concepts, such as “thoughtcrime,” “Newspeak,” and “torture” have entered the English language. Nineteen Eighty-Four is generally considered to be George Orwell's masterpiece.

**Author:** George Orwell

**Country:** United Kingdom

**Language:** English

**Genre(s):** Dystopian, Political novel, Social science fiction

**Publisher:** Secker and Warburg (London)

**Publication date:** 8 June 1949

**Media type:** Print (Hardcover & Paperback) e-book, audio-CD

**Pages:** 326 pp (Paperback edition)

**ISBN:** 978-0-436-20540-0

---

Master thesis by Dustin Lange
Now PhD student at HPI
Topic: similarity search
Occurrence of values in article text: 12 most frequent attributes in infobox_book

72.0 % of the book articles specifying a series in the infobox also contain the series in the article text. 8.7 % of these occurrences could only be found by separately searching for parts of these values.
On average, 42.2% of the infobox_album attribute values can be found in the article text. 38.2% of these occurrences could only be found by separately searching for parts of these values.
Architecture of iPopulator

1. Input Handling
   - Wikipedia Raw Input File
   - Article Text
   - Attribute-Value Pairs

2. Structure Analysis
   - Article
   - Attribute Value Patterns

3. Construction of Training and Test Data
   - Training Data
   - Test Data

4. Attribute Value Extraction
   - Extracted Attribute-Value Pairs
   - Attribute

5. Evaluator
   - Evaluation Results
   - Article
   - Extracted Attribute-Value Pairs

Felix Naumann | Cleansing Web Data | BEWEB 2011
Structure Analysis

- Values of an attribute often share similar structure.
  - Extract value parts
  - Constructing homogeneous values from parts

- Determine common structure for each infobox template attribute
- Example: number_of_employees from infobox_company
Training Data and Extraction

- Exploit existing infoboxes as training data
- Mark occurrences of infobox attribute values as training examples
  - Similarity measure to label fuzzy occurrences
- Automatic extraction method learns to recognize these occurrences by analyzing token (word-level) features
- Create extractors for thousands of infobox template attributes
- Extract parts of attribute values from different article text positions
- Arrange extracted value parts

Example: IBM employed 399,409 people in 2009.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Number</th>
<th>&quot;(&quot;</th>
<th>Number</th>
<th>&quot;)&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracted value</td>
<td>399,409</td>
<td></td>
<td>2009</td>
<td></td>
</tr>
<tr>
<td>parts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Result value</td>
<td>399,409</td>
<td>(2009)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Evaluation: infobox_planet

22032 Mikekoop
From Wikipedia, the free encyclopedia

22032 Mikekoop (provisional designation: 1999 XB₁₆₄) is a main-belt minor planet. It was discovered through the Lowell Observatory Near-Earth-Object Search at the Anderson Mesa Station in Coconino County, Arizona, on December 9, 1999. It is named after Michael Walter Koop, an American electric engineer and amateur astronomer.

See also:

See also:

F-measure

mp_name alt_names discovered name discovery_site epoch discoverer mp_category dimensions

Attribute

F (part) F (complete) Occurrence rate (part)
Evaluation: infobox_book

![Graph showing F-measure and occurrence rate for various attributes]

- **Attribute**: author, name, series, release_date, genre, pub_date, illustrator, country, publisher, isbn, subject, title_orig

- **F-measure / occurrence rate**
  - F (part)
  - F (complete)
  - Occurrence rate (part)
Evaluation: All Infobox-Templates

Precision / Recall / F-measure

Infobox number

F (part)  P (part)  R (part)
Evaluation on all attributes (>4000) of all infobox templates (>800)

Precision

Attributes ordered by precision
Evaluation on all attributes with Precision > 0.75
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator

- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD

- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration

- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
Challenges: Heterogeneity at all levels

- Source
  - Formats
  - Domain
  - Bandwidth
- Schema
  - Structure
  - Semantics
- Data
  - Formatting
  - Duplicates
- File converters
- Clustering, rules
- Patience
- Schema Mapping
- Domain knowledge
- Scrubbing
- Entity Matching

Now: Examples for each
The problem – a format mess

Commitment position key: SI2.514875.1

<table>
<thead>
<tr>
<th>Year</th>
<th>Amount €</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>99,965,021,40</td>
</tr>
</tbody>
</table>

Subject of grant or contract: 2007-EU-50010-P EasyWay * - K(2008) 8479

Responsible Department: Trans-European Transport Network Executive Agency
Budget line name and number: Financial support for projects of common interest in the trans-European transport network (06.03.03)

Programme: TEN Transport
Co-financing rate: 100,00 %

Beneficiary

Name: ANONYMI ETAIRES EKMETALLEPSIS KAIDAIACHEIRISIS ELLINIKON AKTIKODROMON*TEO AE: SOCIETE ANONYME OF HELLENIC MOTORWAYS
Address: 14342 ATHINA, VITNIS STREET 14-18
Country / Territory: Greece

Name: BUNDESREPUBLIK DEUTSCHLAND*REPUBLIQUE FEDERALE D ALLEMAGNE FEDERAL REPUBLIC OF GERMANY
Address: Country / Territory: Germany

Name: CESKA REPUBLIKA*REPUBLIC OF CZECHIA
Address:

```json
{
    "id": "euFinance#28994",
    "year": 2008,
    "nameOfBeneficiary": "ROBERT BOSCH GMBH*",
    "coordinator": false,
    "countryTerritory": "Germany 70049 STUTTGART",
    "coFinancingRate": "67,51 %",
    "amount": 3199959.00,
    "commitmentPositionKey": "F13.A22622.1",
    "subjectOfGrantOrContract": "MULTISPECTRAL TERAHERTZ, INFRARED ...
    "responsibleDepartment": "Information Society and Media",
    "budgetLineNameAndNumber": "Support for research ..."
}
```
The problem – a domain mess 2008

- What is a company?
  - Def. 1: Entities having a companyName
    - 14,292 companies
  - Def. 2: Entities in a category that starts with 'compan%'
    - 21,753
  - Def. 3: Entities having a wikiPageUsesTemplate with value Template:infobox_company
    - 15,491
The problem – a domain mess 2011

- What is a company? 35,588 candidates
  - Def. 1: Entities having a %companyName%
    - 22,890
  - Def. 2: “Company” according to DBpedia ontology
    - 34,567
  - Def. 3: Entities having a wikiPageUsesTemplate with value %compan%
    - 30,702
Company Template

{{Infobox Company
| name            = The Corporation Company
| logo            = [[Image:Example.png|160px]]
| type            = [[Public company|Public]] ((nyse|TCC1)), ((tyo|TCC1))
| genre           = Corporate histories
| predecessor      = The Wikitory Company
| foundation      = [[New York City]], [[United States|U.S.]]) ((Start date|1900)))
| founder         = Wikip Wikiad
| location_city   = [[Seattle]], [[Washington]]
| location_country = [[United States|U.S.]]
| location        =
| locations       = 300 stores (2000) at [[2000-12-31]]
| area_served     = [[North America]]
| key_people      = Wikip Wikiad <small>[[Entrepreneur|Founder]]</small> <br />
| industries      = Wikip Wikiad <small>[[Chief executive officer|CEO]]</small> <br />
| industry        = [[Publishing]]
| products        = [[Book]]s, [[magazine]]s
| services        = Literary restoration, literary archiving
| revenue         = US$500,000,000 (2000), ((increase)) 5% from 1999
| operating_income = US$350,000,000 (2000) ((steady)) from 1999
| net_income       = US$50,000,000 (2000) ((decrease)) 12% from 1999
| assets          = US$1,500,000,000 at [[2000-12-31]] ((decrease)) 5% from year earlier
| equity          = US$950,000,000 at [[2000-12-31]] ((increase)) 6% from year earlier
| owner           = Wikip Wikiad
| num_employees   = 500 (2000)
| parent          = Mega Corporation Inc.
| divisions       = TCC Company Histories, TCC Magazine Services
| subsid          = Restored Book Company, Super Archives, Ltd.
| footnotes       =
| intl            =
}}
The problem – a schema mess

- Wikipedia/DBpedia: Triples and ill-defined templates invite disaster.

- Schema chaos: Many attribute synonyms
  - Hundreds of different attributes

- 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

- _percent_27_percent_27_percent_27companyName
- _percent_3Cbr/_percent_3ECompanyName
- automatedImagingAssociationCompanyName
- bTcgvuvCompanyName
- bellFoundryCompanyName
- companyNameLocal
- companyNameZh
- companyName_percent_E3_percent_80
- companyName_percent_E3_percent_80_companyName
- companyNameNames
- dvdEuroCompanyName
- europeanTradeAssociationCompanyName
- iceCreamCompanyName
- itIsExpensiveCompanyName
- publicCompanyName
- companyNameEn
- companyNameNamesBigBum
- companyName
Infoboxes with CompanyTemplate

- 1083 different attributes
  - 499 appear only once
- Of the 1083 attributes, 39 distinct ones contain ‘name’ as substring
- 273 companies without any name attribute
34567 companies with 455821 triples
1729 different attributes
  894 appear only once

After cleansing by DBpedia
  34711 companies with 368185 triples
  Only 50 different attributes

- keyPeople 34100
- industry 28720
- foundation 26875
- products 26486
- homepage 25982
- location 24094
- companyName 23297
- companyType 19591
- companyLogo 14644
- numEmployees 11395
- locationCity 9210
- name 8700
- locationCountry 7985
- founder 7867
- revenue 7391
- parent 6468
- type 6358
- areaServed 5842
- logo 5434
- founded 4107
- companySlogan 4053
- netIncome 3528
- genre 3369
- subsid 3288
- headquarters 3191
- airline 2686
- services 2568
- callsign 2391
- icao 2386
- iata 2363
- owner 2303
- fleetSize 2246
- operatingIncome 2246
- hubs 2244
- website 2104
- intl 1996
- defunct 1987
- fate 1944
- slogan 1807
- country 1734
- destinations 1712
- assets 1591
- url 1505
- locations 1384
- divisions 1227
- logoSize 1217
- successor 1211
- distributor 1125
Profiling Companies

![Bar chart showing the distribution of various company attributes such as name, revenue, employees, etc.epochs from 0 to 25000, and attributes like location, products, industry, company name, etc.epochs from left to right.]
<table>
<thead>
<tr>
<th>fieldName</th>
<th>&lt;info&gt;</th>
<th>Dollars Obligated</th>
<th>Current Contract Value</th>
<th>Ultimate Contract Value</th>
<th>Major Agency</th>
<th>Modified Contracting Agency</th>
<th>Contracting Agency</th>
<th>Contracting Office</th>
<th>Program Funding Agency</th>
<th>Program Funding Office</th>
<th>Program Funding Office</th>
<th>Reason for Purchase For DoD</th>
</tr>
</thead>
<tbody>
<tr>
<td>example2</td>
<td></td>
<td>$33,710,000</td>
<td>$33,710,000</td>
<td>$33,710,000</td>
<td>Dept. of Defense</td>
<td>1700: NAVY, Department of the</td>
<td>NAVY, Department of the</td>
<td>N00024</td>
<td>NAVY, Department of the</td>
<td>N00024</td>
<td>N00024</td>
<td>Convenience and Econom</td>
</tr>
<tr>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>map to LegalEntity as recipient</td>
<td></td>
<td></td>
<td>subject = &quot;USSpending&quot;,</td>
</tr>
<tr>
<td>info2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>map to LegalEntity as Parent recipient</td>
<td></td>
<td></td>
<td>amount_curr, amount_ulti</td>
</tr>
</tbody>
</table>
The problem – a data mess

- Poor schemata: No types, no constraints
- 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:

Wal-Mart

Undisclosed

Assets exceed £4 billion GBP


€ bn (as of 2004)
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
Data Profiling

- Continuous profiling
- Extreme heterogeneity
- Incremental
- Understanding
- Profiling
- Approximate
- Interactive

Data & Metadata

Felix Naumann | Cleansing Web Data | BEWEB 2011
Prototype: ProLOD

- Platform for ongoing and future work
  - [https://www.hpi.uni-potsdam.de/naumann/sites/prolod/](https://www.hpi.uni-potsdam.de/naumann/sites/prolod/)

- Steps:
  - Data upload
  - Preprocessing
  - Visualization
ProLOD profiling tasks

- Clustering
  - Hierarchical, based on schema
  - Labeling
- Predicate statistics
  - State-of-the-art profiling for attribute values
  - Value types: literals, internal and external links
  - Data types (String, Text, Integer, Decimal, Date)
  - Strings → determine (normalized) patterns
  - Integers, Decimals → display value ranges
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
Midas – Integration project with IBM Almaden Research Center

- Linked Open Data (Midas, LOD)
  - Integrating DBpedia, Freebase, SEC and FDIC at the level company entities
- Regulatory sources (Midas.Finance)
  - Integrating unstructured/semi-structured data sources containing information about a wide range of entities (e.g., SEC and FDIC)
- Government (Midas.Gov)
  - Integrating structured data from government data sources like usaspending.gov, senate.gov, etc.
  - Persons, legal entities, funding
Five steps for integration

1. Source Selection
2. Schema Matching & Mapping
3. Data Extraction & scrubbing
4. Entity Matching
5. Data Fusion
Five steps – Source selection

- Performed by domain experts
- Criteria
  - Availability and downloadability
  - Coverage of domain (completeness)
  - Complementation with other sources
  - Reputation of source
  - Accuracy of data
  - Cost
  - Other data quality criteria...

Top: Health (57,758)

- Animal (5,432)

- Alternative (4,700)
- Conditions and Diseases (14,289)
- Healthcare Industry (5,652)

- Medicine (10,070)
- Mental Health (4,577)
- Regional (0)

- Addictions (2,302)
- Aging (77)
- Beauty (432)
- Child Health (433)
- Conferences (0)
- Dentistry (533)
- Directories (0)
- Disabilities@ (881)
- Education (165)
- Employment@ (361)
- Environmental Health@ (279)
- Fitness (305)
- History@ (8)
- Home Health (245)
- Insurance@ (131)
- Issues@ (2,003)
- Medical Tourism@ (67)
- Men’s Health (178)
- News and Media (202)
- Nursing (1,109)

- Nutrition (550)
- Occupational Health and Safety (423)
- Organizations (132)
- Pharmacy (2,572)
- Products and Shopping (0)
- Professions (1,537)
- Public Health and Safety (3,064)
- Publications@ (731)
- Reproductive Health (1,812)
- Resources (106)
- Search Engines (11)
- Senior Health (647)
- Senses (297)
- Services (37)
- Specific Substances (581)
- Support Groups (280)
- Teen Health (49)
- Travel Health@ (67)
- Weight Loss (286)
- Women’s Health (513)

dmoz.org
Five steps – Schema matching and schema mapping

- Semi-automated matching
  - Label-based and instance-based

- Challenges:
  - Multi-lingual
  - Homonyms and Synonyms
  - 1:1, 1:n, n:m

- Complex data transformation

<table>
<thead>
<tr>
<th>Final Schema</th>
<th>DBPedia</th>
<th>SEC</th>
<th>Freebase</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpediaURI</td>
<td></td>
<td></td>
<td>/type/object/key</td>
</tr>
<tr>
<td>cik</td>
<td>secCik</td>
<td>CIK</td>
<td></td>
</tr>
<tr>
<td>cik</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRSnumber</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>companyName</td>
<td>companyName, name,</td>
<td>name</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nonProfitName</td>
<td></td>
<td></td>
</tr>
<tr>
<td>address</td>
<td></td>
<td>BusinessAddress, MailingAddress</td>
<td></td>
</tr>
<tr>
<td>locationCity</td>
<td>locationCity, location</td>
<td>BusinessAddress, MailingAddress</td>
<td></td>
</tr>
<tr>
<td>locationCountry</td>
<td>locationCountry, location, showflag</td>
<td>BusinessAddress, MailingAddress</td>
<td></td>
</tr>
<tr>
<td>telephone</td>
<td></td>
<td>BusinessAddress</td>
<td></td>
</tr>
<tr>
<td>symbol</td>
<td></td>
<td>Symbol</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>homepage</td>
<td>homepage, url</td>
<td></td>
<td></td>
</tr>
<tr>
<td>keyPeople (name, title)</td>
<td>keyPeople</td>
<td>KeyPeople</td>
<td></td>
</tr>
<tr>
<td>industry</td>
<td></td>
<td>industry</td>
<td></td>
</tr>
<tr>
<td>products</td>
<td></td>
<td>products, services, genre</td>
<td></td>
</tr>
<tr>
<td>companyType</td>
<td>companyType, type, nonProfitType</td>
<td></td>
<td></td>
</tr>
<tr>
<td>numEmployees</td>
<td>numEmployees, employees</td>
<td></td>
<td></td>
</tr>
<tr>
<td>revenue</td>
<td></td>
<td>revenue</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td></td>
<td>netIncome, grossProfit, earnings, operatingIncome</td>
<td></td>
</tr>
<tr>
<td>foundingYear</td>
<td>foundation, ageProperty</td>
<td></td>
<td>/business/company/founded</td>
</tr>
<tr>
<td>fate</td>
<td>fate, currentStatus, end, dissolved, defunct, successor, origins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>companySlogan</td>
<td>companySlogan, motto, slogan</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Five steps – Data extraction & scrubbing

- Recognize data types
- Regular expressions for multi-valued strings
- Remove spurious values (layout, formatting, ...)
- Standardize formats
- Translate from foreign languages
Five steps – Entity matching

- Duplicate entries
- Linking between entries
- Challenges
  - Fuzzy matching: Similarity measures
  - Data volume: Partitioning algorithms
  - Sparse data
    - “Michael Jordan visited Indianapolis”

**Whoa! Over 100 Results Found**
Five steps – Data fusion

- Combine multiple representations of real-world entities
  - Survivorship, consolidation, etc.
- Resolve data conflicts
  - Conflict resolution functions
  - Reputation / accuracy / freshness -> “truth discovery”

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>0766607194</td>
<td>H. Melville</td>
<td>$3.98</td>
</tr>
<tr>
<td>0766607194</td>
<td>Herman Melville</td>
<td>Moby Dick</td>
</tr>
</tbody>
</table>

- Retain data lineage
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
Multi-Lingual Wikipedia

- **Goal: Schema matching across languages**
  - Complement infobox data
  - Autocomplete for authors
  - Detect errors or inconsistencies
  - Keep values up to date

- **Idea: Use cross-language links across 281 languages**
  (Mar 2011)
Interlanguage links (ILLs)

- First, evaluate quality of ILLs and build duplicate clusters
  - Build connected components using cross-language links (restricted to the six largest languages)
- But, largest weakly connected component has 108 articles
  - 26 English, 26 German, 21 French, 13 Italian, 13 Dutch, 9 Spanish articles
Other large components

Piotr – Peter – Pierre – Stone – Rock – Crag & Tail

Easy Listening – Pop music – World music – Musique folk – Folk – Pueblo - Village

Joint Stock Company – … – Brother
A connected component is **incoherent** if it contains more than one node for any language.

- **Strongly connected components (SCC)**
  - Each node is reachable from each other node
  - 1,067,753 SCCs of which 3,469 are incoherent

- **Bidirectionally connected components (BCC)**
  - Undirected graph of bidirectional components is connected
  - 4,241 BCCs of which 2,980 are incoherent

- **Bi-connected components (2CC)**
  - Each pair of vertices is connected via **two** vertex-independent paths.
  - 8,828 2CCs of which 4,770 are vertex-disjoint

Result: 1,069,948 coherent, connected components
Match schemas of **corresponding** infobox templates only.

Different granularities in templates => n:m mapping

**Idea:** Count co-occurrences of infobox templates in terms of connected components

Apply thresholds:
- **Absolute:** at least 5 co-occurrences
- **Relative:** co-occurrence frequency at least 20% of individual occurrences of the templates

```
Infobox programming language  Infobox Programmiersprache

en:  Infobox software  Infobox Software

Infobox web browser
```
Duplicate-based Schema Matching

- General technique of data is available under both schemas
- Idea: If data coincides for attributes of two schemata, they probably match.

- For each infobox template pair
  - For each article pair
    - For each attribute value pair
      - Determine similarity of values (edit-distance)
      - Store in matrix
  - Aggregate similarities across all articles
  - Perform global matching: bipartite assignment
Duplicate-based Schema Matching

### Coordinates

<table>
<thead>
<tr>
<th>Coordinates:</th>
<th>52°30'2&quot;N 13°23'56&quot;E</th>
</tr>
</thead>
</table>

### Country

<table>
<thead>
<tr>
<th>Country</th>
<th>Germany</th>
</tr>
</thead>
</table>

### Government

- **Governing Mayor**
  - Klaus Wowereit (SPD)
- **Governing parties**
  - SPD / Die Linke
- **Votes in Bundesrat**
  - 4 (of 69)

### Area

- **City**
  - 891.85 km² (344.3 sq mi)
- **Elevation**
  - 34 - 115 m (-343 ft)

### Population (31 March 2010)

- **City**
  - 3,440,441
- **Density**
  - 3,857.6 km² (9,991.3/sq mi)
- **Metro**
  - 4,429,847

### Time zone

- **CET (UTC+1)**
- **CEST (UTC+2)**

### Postal code(s)

<table>
<thead>
<tr>
<th>Postal code(s)</th>
<th>10001–14199</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area code(s)</td>
<td>030</td>
</tr>
<tr>
<td>ISO 3166 code</td>
<td>DE-BE</td>
</tr>
<tr>
<td>Vehicle registration</td>
<td>B</td>
</tr>
</tbody>
</table>

### GDP/Nominal


### NUTS Region

<table>
<thead>
<tr>
<th>NUTS Region</th>
<th>DE3</th>
</tr>
</thead>
</table>

### Website

| Website | berlin.de |

### Basisdaten

<table>
<thead>
<tr>
<th>Basisdaten</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fläche:</strong></td>
</tr>
<tr>
<td><strong>Bevölkerungsdichte:</strong></td>
</tr>
<tr>
<td><strong>BIP:</strong></td>
</tr>
<tr>
<td><strong>Höhe:</strong></td>
</tr>
<tr>
<td><strong>Geografische Lage:</strong></td>
</tr>
<tr>
<td><strong>Zeitzone:</strong></td>
</tr>
<tr>
<td><strong>Postleitzahlen:</strong></td>
</tr>
<tr>
<td><strong>Vorwahl:</strong></td>
</tr>
<tr>
<td><strong>Kfz-Kennzeichen:</strong></td>
</tr>
<tr>
<td><strong>Gemeindeschlüssel:</strong></td>
</tr>
<tr>
<td><strong>ISO 3166-2:</strong></td>
</tr>
<tr>
<td><strong>UN/LOCODE:</strong></td>
</tr>
<tr>
<td><strong>Website:</strong></td>
</tr>
</tbody>
</table>

### Politik

<table>
<thead>
<tr>
<th>Politik</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reg. Bürgermeister:</strong></td>
</tr>
<tr>
<td><strong>Reg. Parteien:</strong></td>
</tr>
<tr>
<td><strong>Sitzverteilung im Abgeordnetenhaus:</strong></td>
</tr>
</tbody>
</table>

Felix Naumann | Cleansing Web Data | BEWEB 2011
### Evaluation

- Qualitative evaluation via hand-crafted attribute mappings
  - 96 infobox template pairs
  - 1,417 expected attribute pairs

<table>
<thead>
<tr>
<th></th>
<th>en</th>
<th>de</th>
<th>fr</th>
<th>en</th>
<th>nl</th>
<th>de</th>
<th>fr</th>
<th>de</th>
<th>nl</th>
<th>fr</th>
<th>nl</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>91.97</td>
<td>92.28</td>
<td>95.15</td>
<td>90.78</td>
<td>91.67</td>
<td>93.85</td>
<td>92.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>94.17</td>
<td>96.83</td>
<td>94.80</td>
<td>92.06</td>
<td>93.22</td>
<td>92.82</td>
<td>94.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>F₁ Score</strong></td>
<td>93.06</td>
<td>94.50</td>
<td>94.97</td>
<td>91.42</td>
<td>92.44</td>
<td>93.33</td>
<td>93.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Felix Naumann | Cleansing Web Data | BEWEB 2011
Overview

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
Motivation – Wealth of Open Gov Data

Biographical Directory of the United States Congress
1774 - Present

European Commission
Financial Transparency System

Welcome to the European Parliament

USA Spending.gov
Government spending at your fingertips

Office of Management and Budget

Felix Naumann | Cleansing Web Data | BEWEB 2011
Companies, Agencies, and People
Interesting queries

- Find all *classmates* of George W. Bush who, during his term, have worked at a company that has received government funding.

- For each member of congress, find all earmarks awarded to organizations that have *employed a relative* of that member of congress.

- For each government employees, find all companies that have received funding supported by that member and have *employed him after/before their term in congress*.

- Goal: Demonstrate the power of
  - *Joins*: Find unknown connections
    `<person – university|company|fund – person>`
  - *Grouping and aggregation*: Combine data about parties, companies, and persons; calculate sums.
  - *Sorting*: Order results by funding amount
  - *Sets*: “for each ... find all ...”
Five steps for integration

1. Source Selection
2. Schema Matching & Mapping
3. Data Extraction & scrubbing
4. Entity Matching
5. Data Fusion
## Data sources so far

<table>
<thead>
<tr>
<th>Source</th>
<th>Num. of entities</th>
<th>Num. of attributes</th>
<th>Format</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>US Spending</td>
<td>1.7m</td>
<td>122</td>
<td>XML</td>
<td>all gov spending</td>
</tr>
<tr>
<td>US Earmarks</td>
<td>20,000</td>
<td>37</td>
<td>CSV</td>
<td>anonymous guarantees</td>
</tr>
<tr>
<td>US Congress</td>
<td>12,000</td>
<td>8</td>
<td>HTML</td>
<td>members of congress since 1744, incl. bio</td>
</tr>
<tr>
<td>DE Party Donations</td>
<td>1,500</td>
<td>4</td>
<td>HTML</td>
<td>Donations &gt; 20,000 €</td>
</tr>
<tr>
<td>EU Finance</td>
<td>122,000</td>
<td>11</td>
<td>HTML</td>
<td>EU spending</td>
</tr>
<tr>
<td>EU Agric. Subventions</td>
<td>207,000</td>
<td>8</td>
<td>HTML</td>
<td>EU spending</td>
</tr>
<tr>
<td>EU Parliam. Data</td>
<td>900</td>
<td>14</td>
<td>HTML</td>
<td>members of parliament</td>
</tr>
<tr>
<td>Freebase Person Data</td>
<td>1.8m</td>
<td>32</td>
<td>TSV</td>
<td>person data</td>
</tr>
</tbody>
</table>
Data – Mapping and Scrubbing

- **fund**
  - recipient
  - sponsor
  - family
  - friends

- **abstract object**
  - receiving and spending money

- **person / politician**
  - employment
  - family
  - friends

- **legal entity**
  - hierarchy

JSON example:
```json
{
  "id": "euFinance#28994",
  "year": 2008,
  "nameOfBeneficiary": "ROBERT BOSCH GMBH",
  "coordinator": false,
  "countryTerritory": "Germany 70049 STUTTGART",
  "coFinancingRate": "67.51 %",
  "amount": 3199959.00,
  "commitmentPositionKey": "F13.A22622.1",
  "subjectOfGrantOrContract": "MULTISPECTRAL TERAHERTZ, INFRARED ...",
  "responsibleDepartment": "Information Society and Media",
  "budgetLineNameAndNumber": "Support for research ...
```
Data – Cleansing

■ Deduplication / Entity Matching
  □ Intra Source Consolidation
  □ Intra Source Duplicate Detection
    ◊ Duplicate Detection Toolkit – DuDe
    ◊ Hundreds of duplicates within original sources
  □ Entity Matching across Sources
    ◊ Augment discovered Person Data with Freebase Info
    ◊ Jaro-Winkler and Monge-Elkan distance

■ Entity Fusion
  ◊ Dempster-Shafer-Theory
Web Data abounds
- Linked, open, and otherwise
- iPopulator

Web Data stinks
- Dirt, grime, and some surprises
- ProLOD – Profiling LOD

Cleansing and Integration
- ...of mops and brooms
- Cross-Language Integration

Government data
- Politicians, friends, and funds
- The GovWILD experience
http://govwild.org

- 200,000 persons
- 248,000 legal entities
- 1,000,000 funds

- Keyword Queries

- Linked Data Interface (dereference URIs)

- Exploration of entities mentioned in New York Times articles

- Data Download (RDF, SQL Dump, JSON files)
Barack Obama
Barack Hussein Obama II (born in 1961) is the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as a United States Senator from Illinois, from January 2005 until he resigned after his election to the presidency in November 2008. A native of Honolulu, Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at the University of Chicago Law School from 1992 to 2004. Obama served three terms in the Illinois Senate from 1997 to 2004. Following an unsuccessful bid against a Democratic incumbent for a seat in the U.S. House of Representatives in 2000, he ran for United States Senate in 2004.[1] Several events brought him to national attention during the campaign, including his victory in the March 2004 Democratic primary and his keynote address at the Democratic National Convention in July 2004. He won election to the U.S. Senate in November 2004. His presidential campaign began in February 2007, and after a close campaign in the 2008 Democratic Party presidential primaries against Hillary Rodham Clinton, he won his party's nomination. In the 2008 general election, he defeated Republican nominee John McCain and was inaugurated as president on January 20, 2009.4

Earmarks
<table>
<thead>
<tr>
<th>Predicate</th>
<th>Value</th>
<th>Info</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontology:placeOfBirth</td>
<td>Honolulu, Hawaii</td>
<td></td>
</tr>
<tr>
<td>ontology:birthYear</td>
<td>1961</td>
<td></td>
</tr>
<tr>
<td>ontology:lastName</td>
<td>Obama</td>
<td></td>
</tr>
<tr>
<td>rdf:type</td>
<td>ontology:Person</td>
<td></td>
</tr>
<tr>
<td>ontology:employment</td>
<td>Democratic Party party 269</td>
<td></td>
</tr>
<tr>
<td>rdfs:label</td>
<td>Barack Obama</td>
<td></td>
</tr>
<tr>
<td>ontology:worksFor</td>
<td>Democratic Party</td>
<td></td>
</tr>
<tr>
<td>ontology:birthDay</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>ontology:birthMonth</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
<tr>
<td>ontology:sponsorOf</td>
<td>UsEarmark</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- Web Data abounds
  - Linked, open, and otherwise
  - iPopulator
- Web Data stinks
  - Dirt, grime, and some surprises
  - ProLOD – Profiling LOD
- Cleansing and Integration
  - ...of mops and brooms
  - Cross-Language Integration
- Government data
  - Politicians, friends, and funds
  - The GovWILD experience
References

- **Extracting Structured Information from Wikipedia Articles to Populate Infoboxes**
  Dustin Lange, Christoph Böhm, and Felix Naumann
  *Proceedings of the 19th Conference on Information and Knowledge Management (CIKM) 2010, Toronto, Canada (Extended version available as technical report)*

- **Profiling Linked Open Data with ProLOD**
  Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, David Sonnabend
  Workshop *New Trends in Information Integration* (NTII) 2010, Long Beach, USA

- **Linking Open Government Data: What Journalists Wish They Had Known**
  Christoph Böhm, Felix Naumann, Markus Freitag, Stefan George, Norman Höfler, Martin Köppelmann, Claudia Lehmann, Andrina Mascher, and Tobias Schmidt.
  *Honorable Mention* at Linked Data Triplification Challenge 2010 @ I-Semantics, Graz. (link to GovWILD)

- **DuDe: The Duplicate Detection Toolkit**
  Uwe Draisbach and Felix Naumann: QDB 2010 Workshop at VLDB, Singapore