Extreme Web Data Integration

August 14, 2010
Felix Naumann
Acknowledgements

- @IBM Almaden
  - Howard Ho, Mauricio Hernandez, Rajasekar Krishnamurthy, Lucian Popa, Roxana Stanoi

- @HPI
  - Christoph Böhm
  - Bachelor project student team

- And elsewhere
  - Antonio Sala (University of Modena)
  - Chis Bizer (dbpedia)
  - Open Data community
Overview

- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds
### DBpedia - Extraction

```json
{{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 the Institute of Radio Engineers
 | key_people = Mr. Pedro A. Ray, Current President
 | area_served = Worldwide
 | focus = Electrical, Electronics, and Information Technology [http://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 =
}}
```

### IEEE

**Type**: Professional Organization  
**Founded**: January 1, 1963  
**Origins**: Merger of the American Institute of Electrical Engineers and the Institute of Radio Engineers  
**Key people**: Mr. Pedro A. Ray, Current President  
**Area served**: Worldwide  
**Focus**: Electrical, Electronics, and Information Technology  
**Method**: Industry standards, Conferences, Publications  
**Revenue**: US$330 million  
**Members**: 395,000+  
**Website**: [www.ieee.org](http://www.ieee.org)
DBpedia statistics

1. Core Datasets

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

- **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**

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

...
Use cases

- General purpose integration: Create rich knowledge bases
  - Semantic Web
  - Improved Search
  - Link creation
  - Cleansing
- Domain specific integration
  - Creation of high quality data sets: Complete & accurate
  - Enhancement of organization-internal data
  - Create reference data sets
Two Flavors of Integration

- **Web Data Integration**
  - Structured/semi-structured data
  - Exposed by different data providers
  - Steps: Source selection, data extraction from individual fields/attributes, scrubbing, entity matching, data transformation, data fusion

- **Unstructured Entity Integration**
  - Unstructured data
  - Small set of focused data sources.
  - Steps: Information extraction (from large text/html documents), Entity Resolution and Data Mapping / Fusion.
Midas – Integration project with IBM Almaden Research Center

- Linked Open Data (Midas, LOD)
  - Integrating DBpedia, Freebase, SEC and FDIC at the level company entities

- Government (Midas.Gov)
  - Integrating structured data from government data sources like usaspending.gov, senate.gov, etc.
  - Persons, legal entities, funding

- Regulatory sources (Midas.Finance)
  - Integrating unstructured/semi-structured data sources containing information about a wide range of entities (e.g., SEC and FDIC)
Overview

- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds
Challenges: Heterogeneity at all levels

- **Source**
  - Formats
  - Domain
  - Bandwidth
  - File converters
  - Clustering, rules
  - Patience

- **Schema**
  - Structure
  - Semantics
  - Schema Mapping
  - Domain knowledge

- **Data**
  - Formatting
  - Duplicates
  - Scrubbing
  - Entity Matching
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 ETAIREIA EKMETALLEFISIS KAIIDACHAIRISIS ELLINIKON AFTOKINITODROMON*TEO AE SOCIETE ANONYME OF HELLENIC MOTORWAYS
Address: 14342 ATHINA, VITNIS STREET 14-18 Country / Territory: Greece

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

Name: CESKA REPUBLIKA*REPUBLIC OF CZECH REPUBLIC
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

What is a company?

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

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

**Def. 3:** Entities having a `wikiPageUsesTemplate` with value `Template:infobox_company`
- 15491

Felix Naumann | Extreme Web Data Integration | NFIC 2010
The problem – a schema mess

- Wikipedia/DBpedia: Triples and ill-defined templates invite disaster.
- 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
Profiling 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
Infoboxes with CompanyTemplate

- 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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>20617</td>
<td></td>
</tr>
<tr>
<td>products</td>
<td>18176</td>
<td></td>
</tr>
<tr>
<td>wikiPageUsesTemplate</td>
<td>18048</td>
<td></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>companyName</td>
<td>13355</td>
<td></td>
</tr>
<tr>
<td>companyLogo</td>
<td>9006</td>
<td></td>
</tr>
<tr>
<td>numEmployees</td>
<td>6207</td>
<td></td>
</tr>
<tr>
<td>revenue</td>
<td>5030</td>
<td></td>
</tr>
<tr>
<td>locationCity</td>
<td>4098</td>
<td></td>
</tr>
<tr>
<td>locationCountry</td>
<td>3212</td>
<td></td>
</tr>
<tr>
<td>companySlogan</td>
<td>2815</td>
<td></td>
</tr>
<tr>
<td>areaServed</td>
<td>2557</td>
<td></td>
</tr>
<tr>
<td>relatedInstance</td>
<td>2284</td>
<td></td>
</tr>
<tr>
<td>type</td>
<td>2152</td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>2054</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>2036</td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>1663</td>
<td></td>
</tr>
<tr>
<td>founder</td>
<td>1597</td>
<td></td>
</tr>
<tr>
<td>subsid</td>
<td>1232</td>
<td></td>
</tr>
<tr>
<td>nihongoProperty</td>
<td>1141</td>
<td></td>
</tr>
<tr>
<td>slogan</td>
<td>1087</td>
<td></td>
</tr>
<tr>
<td>coorTitleDmsProperty</td>
<td>960</td>
<td></td>
</tr>
<tr>
<td>logo</td>
<td>925</td>
<td></td>
</tr>
<tr>
<td>services</td>
<td>904</td>
<td></td>
</tr>
<tr>
<td>operatingIncome</td>
<td>896</td>
<td></td>
</tr>
<tr>
<td>owner</td>
<td>680</td>
<td></td>
</tr>
<tr>
<td>otheruses4Property</td>
<td>510</td>
<td></td>
</tr>
<tr>
<td>intl</td>
<td>503</td>
<td></td>
</tr>
<tr>
<td>forProperty</td>
<td>467</td>
<td></td>
</tr>
<tr>
<td>divisions</td>
<td>429</td>
<td></td>
</tr>
<tr>
<td>date</td>
<td>422</td>
<td></td>
</tr>
<tr>
<td>locations</td>
<td>419</td>
<td></td>
</tr>
<tr>
<td>companyName</td>
<td>13355</td>
<td></td>
</tr>
<tr>
<td>name</td>
<td>2036</td>
<td></td>
</tr>
<tr>
<td>surname</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>railroadName</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>companyNickname</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>pastNames</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>absNameProperty</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>drvNameProperty</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>labelName</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>logoFilename</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>dvdEuroCompanyName</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>filename</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>longName</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>websitename</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>alternativeNames</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>birthname</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>brandName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>bTcgvvuvCompanyName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>companyNameLocal</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>companyNamesBigBum</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>europeanTradeAssociation CompanyName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>familyCorporationCompanyName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>formerNames</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>fukCompanyName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>golfFacilityName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>hangulName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>iceCreamCompanyName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>nativeName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>nickname</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>officialName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>oldName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>organisationName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>publicCompanyName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>renamed</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>shortName</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>wineeryName</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
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 = Wiped 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 = Wiped Wikiad <small>[[Entrepreneur|Founder]]</small> <br />
  | industry = [[Publishing]]
  | products = [[Book]]s, [[magazine]]s
  | services = Literary restoration, literary archiving
  | revenue = US$50,000,000 (2000), ((increase)) 5% from 1999
  | operating_income = US$350,000,000 (2000) ((steady)) from 1999
  | net_income = US$500,000 (2000) ((decrease)) 1% from 1999
  | assets = US$1,500,000,000 at [[2000-12-31]] ((decrease)) 4% from year earlier
  | equity = US$950,000,000 at [[2000-12-31]] ((increase)) 6% from year earlier
  | owner = Wiped Wikiad
  | num_employees = 1,500 (2000)
  | parent = Mega Corporation Inc.
  | divisions = TCC Company Histories, TCC Magazine Services
  | subsid = Restored Book Company, Super Archives, Ltd.
  | footnotes =
  | intl =
)
<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>Reason For Purchasing Agency</th>
<th>Federal Funding Office</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>example1</strong></td>
<td></td>
<td>$220,989,132</td>
<td>$220,989,132</td>
<td>$220,989,132</td>
<td>Dept. of Defense</td>
<td>Defense Logistics Agency</td>
<td>Navy, Department of the Navy</td>
<td>SP0000</td>
<td>Defense Logistics Agency</td>
<td>SP0000</td>
<td>Invalid code</td>
<td></td>
</tr>
<tr>
<td><strong>example2</strong></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 Navy</td>
<td>NAVY, Department of the Navy</td>
<td>N00024</td>
<td>NAVY, Department of the Navy</td>
<td>N00024</td>
<td>Convenience and Economy</td>
<td></td>
</tr>
<tr>
<td><strong>info</strong></td>
<td></td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
</tr>
<tr>
<td><strong>info2</strong></td>
<td></td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
<td>never null</td>
</tr>
<tr>
<td><strong>scrubbing</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>
</tbody>
</table>

Phew!
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


Image: green_up.png

€ bn (as of 2004)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fieldName</td>
<td>example</td>
<td>type</td>
<td>info</td>
<td>not relevant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>id</td>
<td>usEarmarks#123_P</td>
<td>string</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>firstName</td>
<td>string</td>
<td>string</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>middleName</td>
<td>string</td>
<td>string</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>lastName</td>
<td>string</td>
<td>string</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>nameAddition</td>
<td>&quot;(Jr.)&quot;</td>
<td>[String] Dictionary</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>originals</td>
<td>[&quot;fileName=876867&quot;, ...]</td>
<td>[string] ID only</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>addresses</td>
<td>[(&quot;string&quot;)]</td>
<td>[record]</td>
<td>info</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>street</td>
<td>&quot;5th Avenue, 1&quot;</td>
<td>string</td>
<td>other countries?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>additionalLine</td>
<td>&quot;Post office: 123&quot;</td>
<td>string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>city</td>
<td>&quot;New York&quot;</td>
<td>string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>zipCode</td>
<td>&quot;012G4&quot;</td>
<td>string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>state</td>
<td>&quot;New York&quot;</td>
<td>string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>district</td>
<td>2 int</td>
<td>int</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>country</td>
<td>&quot;United States&quot;</td>
<td>string</td>
<td>so.org/iso/en</td>
<td>United States</td>
<td>US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td>ish_country</td>
<td>country</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>birth: {year: month: day:}</td>
<td>record with int</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>death: {year: month: day:}</td>
<td>record with int</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>gender</td>
<td>&quot;male&quot;, &quot;female&quot;</td>
<td>string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>placeOfBirth</td>
<td>string</td>
<td>bio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>placeOfDeath</td>
<td>string</td>
<td>bio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>nationalities</td>
<td>[France]</td>
<td>[string]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>religion</td>
<td><a href="http://www.vitobonsignor">http://www.vitobonsignor</a></td>
<td>string</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>websites</td>
<td>[string]</td>
<td>[string]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>biography</td>
<td>text</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>bibliography</td>
<td>text</td>
<td>text</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>ethnicity</td>
<td>White</td>
<td>string</td>
<td>politician</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>professions</td>
<td>[&quot;politician&quot;]</td>
<td>[string]</td>
<td>politician</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bereit | Sheet1 | 4 |
Overview

- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds
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@ (131)
- 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>companyName</td>
<td>companyName, name, nonProfitName</td>
<td>name</td>
<td>/type/object/name, /common/ /location/mailing_address/city</td>
</tr>
<tr>
<td>address</td>
<td>BusinessAddress, MailingAddress</td>
<td>/location/mailing_address/city</td>
<td></td>
</tr>
<tr>
<td>locationCity</td>
<td>locationCity, location</td>
<td>BusinessAddress, MailingAddress</td>
<td>/location/mailing_address/city</td>
</tr>
<tr>
<td>locationCountry</td>
<td>locationCountry, location, showflag</td>
<td>BusinessAddress, MailingAddress</td>
<td>/location/mailing_address/city</td>
</tr>
<tr>
<td>telephone</td>
<td>BusinessAddress</td>
<td></td>
<td></td>
</tr>
<tr>
<td>symbol</td>
<td>symbol</td>
<td>Symbol</td>
<td>/business/company/ticker_sym</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>/business/employer/employee</td>
</tr>
<tr>
<td>industry</td>
<td>industry</td>
<td></td>
<td>industry</td>
</tr>
<tr>
<td>products</td>
<td>products, services, genre</td>
<td></td>
<td></td>
</tr>
<tr>
<td>companyType</td>
<td>companyType, type, nonProfitType</td>
<td>company_type</td>
<td></td>
</tr>
<tr>
<td>numEmployees</td>
<td>numEmployees, employees</td>
<td></td>
<td></td>
</tr>
<tr>
<td>revenue</td>
<td>revenue</td>
<td></td>
<td></td>
</tr>
<tr>
<td>netIncome</td>
<td>netIncome, grossProfit, earnings, operatingIncome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>foundingYear</td>
<td>foundation, ageProperty</td>
<td>/business/company/founded</td>
<td></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>subsidiary</td>
<td>subsidiary</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Felix Naumann | Extreme Web Data Integration | NFIC 2010
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”

Find People

Find People | Find a Business

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>City, State or ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Michael</td>
<td>Jordan</td>
<td>CA</td>
</tr>
</tbody>
</table>

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”
- Retain data lineage

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Price</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>0766607194</td>
<td>H. Melville</td>
<td>$3.98</td>
<td></td>
</tr>
<tr>
<td>0766607194</td>
<td>Herman Melville</td>
<td>$5.99</td>
<td></td>
</tr>
</tbody>
</table>
Web Data abounds
  - linked, open, and otherwise

Web Data stinks
  - dirt, grime, and some surprises

Cleansing and Integration
  - of mops and brooms

The GovWILD experience
  - politicians, friends, and funds
Motivation – Wealth of Open Gov Data

- Biographical Directory of the United States Congress
- European Commission
- Financial Transparency System
- USA Spending.gov
- Office of Management and Budget
- Unklarheiten.de
Interesting queries

- Find all classmates of George W. Bush who, during Bush’s 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 member of congress, find all companies that have received funding supported by that member and have employed him/her after their term in congress.

Goal: Demonstrate the power of

- **Sets**: Instead of researching individuals, write queries against large sets of persons
- **Joins**: Make unknown connections, for instance connecting persons through their universities or connecting persons with companies in multiple ways (employment and funding)
- **Grouping and aggregation**: Combine information about parties, companies, and persons and find averages and sums.
- **Sorting**: Order results by funding amount to find top results.
## 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

- Sponsor
- Fund
- Recipient
- Abstract object: receiving and spending money
- Person / Politician
- Employment
- Legal entity
- Hierarchy

Example JSON:
```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 ..."
}
```
legal entity: {
    "_id": "euFinance#28994_L1",
    "addresses": [  
        { "country": "Germany", 
          "zipCode": "70049",
          "city": "Stuttgart" } ],
    "name": "Robert Bosch",
    "originals": [ "euFinance#28994", "euFinance#37025" ],
    "receivedFunds": [ "euFinance#28994_F", "euFinance#37025_F" ],
    "type": { "form": "GmbH",
              "category": "company" }
}

fund: {
    "_id": "euFinance#28994_F",
    "amount": 3199959,
    "currency": "EUR",
    "date": { "year": 2008 },
    "originals": [ "euFinance#28994" ],
    "recipients": [ "euFinance#28994_L1" ],
    "sponsors": [ "euFinance#42090_L2" ]
}

"subjectOfGrantOrContract": "MULTISPECTRAL TERAHERTZ, INFRARED ...",
"responsibleDepartment": "Information Society and Media",
"budgetLineNameAndNumber": "Support for research ..."
Data – Cleansing

- Deduplication
  - 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
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)
Summary

- Web Data abounds
  - linked, open, and otherwise
- Web Data stinks
  - dirt, grime, and some surprises
- Cleansing and Integration
  - of mops and brooms
- The GovWILD experience
  - politicians, friends, and funds