# Linking Open Government Data: What Journalists Wish They Had Known

Christoph Böhm[1], Felix Naumann[1], Markus Freitag[2], Stefan George[2], Norman Höfler[2], Martin Köppelmann[2], Claudia Lehmann[2], Andrina Mascher[2], Tobias Schmidt[2]
Hasso Plattner Institute, Potsdam, Germany
[1] firstname.lastname@hpi.uni-potsdam.de [2] firstname.lastname@student.hpi.uni-potsdam.de

## ABSTRACT

Many government organizations publish a variety of data on the web to facilitate transparency. The multitude of sources has resulted in heterogeneous structures and formats as well as varying quality of such data. We report on a project dubbed *GovWild* (Government Web Data Integration for Linked Data) that integrates and cleanses open government data at a large scale. Also, we point to the unified and clean integration result, published as Linked Open Data at govwild.hpi-web.de, and feature our web application to showcase the usability of the created dataset.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Misc.

## General Terms

Semantics

## Keywords

Linked Open Data, Open Government Data, NYT Data, Integration

## 1. INTEGRATING GOVERNMENT DATA

A particularly interesting domain among the new movement of disclosing data openly on the web is that of government data. In theory, any citizen can monitor the actions of elected officials and government agencies. In practice however, gaining access to such raw data, placing it into a meaningful context, and extracting useful information is extremely difficult: Data is provided in a variety of formats and schemas, it is erroneous, entities are not characterized by globally consistent IDs, and execution of ad-hoc analysis-type user queries on such vast amounts of data is difficult to achieve efficiently. To make proper use of it, an end-to-end integration solution is needed.

In this paper we report on a project that integrates vastly different types of open government data into a concise, clean, and well-structured dataset. In particular we address the problems of data extraction, scrubbing, transformation, entity matching and data fusion. It is based on previous work in IBM's Midas project [6]. Also, we discuss the generation of Linked Open Data (LOD) from the integration result. Specifically, we have integrated several US and EU public sector data sources and have interlinked them with the NYT dataset to showcase the usefulness of such integrated data. Our first contribution is the resulting dataset ($\approx 18.6$ million triples), which allows insightful queries against the network of politicians, companies, and government spending, e.g.:

- *Find all graduates from the university President X went to who, during X's term, have worked at a company that has received government funding, ordered by amount.*
- *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.*

This data can be easily combined with other linked open data and thus forms an addition to the linked data cloud. Second, on top of this dataset we have built a Web-application to browse, query, and visualize this data. For instance, it allows the discovery of interesting connections among entities mentioned in NYT articles.

## 2. DATA SOURCES

Integrated data sources originate from US and EU government agencies. We chose those two geographically distinct origins, because there is a major effort in these regions to publish government data. In addition, it forced us to define and implement a universal approach to capture the data suitable for an end-to-end integration process. We focused on three simple entity types: *person* (politicians, company key-persons, etc.), *legal entity* (companies, government agencies, parties), and *fund* (grants, subventions, earmarks, donations, etc.).

The selection of the sources to be integrated in this project was driven by several factors: First, we targeted official sources that allow interesting combinations of previously separated information. With 'interesting' we mean facilitating the discovery of new relations, e.g., the CEO of a company that received a funding sponsored by some other person that has worked for the same company in the past. Also, we want to find 'official confirmations' of existing rela-

*I-SEMANTICS* 2010 Graz, Austria

tions among entities, e.g., a politician and a company mentioned in the same newspaper article. Finally, we favored sources that appear to be complete (in the sense of few null values). To achieve timely results, we usually used data from the most recent available year.

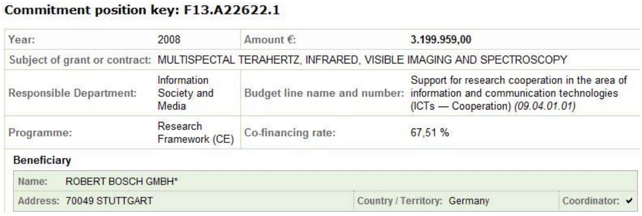
Table 1 lists all integrated data sources and respective sizes as well as source formats. For the US we integrated spending-, earmarks-, and congress-data. The first source contains all spending related to all federal contracts, e.g., military expenditures. The earmarks dataset covers anonymously authored guarantees of federal funds to particular recipients – however, the congress member who sponsors it is known. The congress dataset contains members of the US congress from 1774 until present, including biographical information, such as education and family relationships. For the European Union the finance data (plus agriculture subventions for Germany) can be considered an equivalent to the US spending data. The EU parliament data corresponds to the US congress data but is more detailed. For the sake of interestingness, we also include a dataset that covers German party donations. Additionally, we used Freebase data to augment information found in other sources and to link our dataset to the LOD cloud. Based on these links we have created an illustrative use case – the evaluation and discovery of connections among entities mentioned in NYT articles (see Sec. 5).

| data source | num. of entities | num. of attributes | format |
|---|---:|---:|---:|
| US Spending[1] | 1,724,655 | 122 | XML |
| US Earmarks[2] | 19,753 | 37 | CSV |
| US Congress[3] | 12,470 | 8 | HTML |
| DE Party Donations[4] | 1,521 | 4 | HTML |
| EU Finance[5] | 121,495 | 11 | HTML |
| EU Agric. Subventions[6] | 207,304 | 8 | HTML |
| EU Parliament Data[7] | 904 | 14 | HTML |
| Freebase Data[8] | 1,780,821 | 32 | TSV |

**Table 1: Data sources under consideration.**

Table 1 also states the different source formats of the data. Note that most of the data comes as (unstructured) HTML and we thus developed a set of site-specific crawlers. Figure 1 gives an example of such data from `ec.europa.eu/beneficiaries/fts`. A further real-world challenge to deal with is the often poor site availability. Moreover, some data only appears as raw text on the web, e.g. biographic information. Here, we leveraged SystemT [1] to perform extraction of structured information. After collecting the data, we transformed it into a generic JSON format as depicted in Fig. 2. This format is the basis of the integration process that we explain in the following section.

## 3. INTEGRATION FLOW



**Commitment position key: F13.A22622.1**

| | | | |
|---|---|---|---|
| Year: | 2008 | Amount €: | 3.199.959,00 |
| Subject of grant or contract: | MULTISPECTAL TERAHERTZ, INFRARED, VISIBLE IMAGING AND SPECTROSCOPY | | |
| Responsible Department: | Information Society and Media | Budget line name and number: | Support for research cooperation in the area of information and communication technologies (ICTs — Cooperation) *(09.04.01.01)* |
| Programme: | Research Framework (CE) | Co-financing rate: | 67,51 % |

**Beneficiary**

| | | | |
|---|---|---|---|
| Name: ROBERT BOSCH GMBH* | | | |
| Address: 70049 STUTTGART | Country / Territory: Germany | Coordinator: ✓ | |

**Figure 1: Example HTML data from ec.europa.eu**

```
{ "_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" : "MULTISPECTAL TERAHERTZ, INFRARED ...
  "responsibleDepartment" : "Information Society and Media",
  "budgetLineNameAndNumber" : "Support for research ...
}
```

**Figure 2: Raw data from Fig. 1 in JSON format.**

Given raw JSON data like that shown in Fig. 2 we performed an end-to-end integration process to create a clean, duplicate-free, consistent, and well-structured dataset from the heterogeneous input. Note that what we describe in the following is a generic workflow and therefore allows the incorporation of more data sources with little effort. However, one would, of course, have to create source specific extraction techniques. The integration flow comprises the following steps:

Mapping and Scrubbing: In the first phase we map attributes from the sources to a simple global schema covering *persons*, *legal entities*, and *funds*. This step includes scrubbing, i.e., cleansing on data value level. Here, for instance, we decompose names and standardize date formats.

Data Transformation: The source data structure usually does not entirely match our global schema, e.g., there is a single entry that contains organization as well as funding information. The transformation step separates data of different types. Previously, we have mapped attributes. Now we transform the JSON objects as a whole, which results in three sets of objects, each specific to an entity type. Figure 3 shows an example for a legal entity and a fund (mostly resulting from the source in Fig.1 and Fig.2; more information will be added in the Fusion step).

Deduplication: The next step identifies intra-source duplicates. Besides real duplicates, e.g., a few hundreds within Freebase person data, we resolve initial entries (say a line in a tsv file – now a JSON object) that, together with multiple other entries, represent a single real-world object. This happens when entity information is spread across a source.

Subsequently, we perform entity matching across data sources using our Duplicate Detection Toolkit (DuDe) [3]. For instance, we identify persons from Freebase that match a given person from the US congress data. Such a match leads to an augmentation of the initial congress data in the next step. However, detecting such matches is not a trivial task, because we do not have clean Linked Open Data yet. For instance, for person's names we leverage the Jaro-Winkler [9] and Monge-Elkan [5] distance measures as well as

[1] http://www.usaspending.gov/fpds
[2] http://max.omb.gov/earmarks-public
[3] http://bioguide.congress.gov
[4] http://www.parteispenden.unklarheiten.de
[5] http://ec.europa.eu/beneficiaries/fts
[6] http://www.agrar-fischerei-zahlungen.de
[7] http://www.europarl.europa.eu
[8] http://download.freebase.com/datadumps

look-up-lists for nicknames. Additionally, we apply domain-specific knowledge: For instance, a person $A$ that issued an earmark cannot match a person $B$ from the Congress data, if the earmark was issued in a period different from the time when $B$ was a member of congress.

Entity Fusion: Last, we fuse matched entities to obtain a single representation. We apply Dempster-Shafer-Theory [7] to induce good weights for the attribute value selection based on data quality scores. These (source-specific) scores were empirically determined and combine information such as whether a source is official or not, and results from certain quality tests.

The integration result is a set of entity-type-specific JSON objects representing distinct real-world objects. Figure 3 gives an example of a legal entity and a fund. Note that we retain data lineage information, i.e., the *originals* attribute, and at this point we have a set of interlinked objects, e.g., the *receivedFunds*, *recipients*, *sponsors* attributes contain links to other objects.

```
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" ]
}
```

**Figure 3: JSON object after Data Transformation and Entity Fusion.**

## 4. TRIPLIFICATION

The integration process yields three sets of entities, i.e., persons, legal entities, and funds. For the conversion to RDF triples we store them in a traditional RDBMS and leverage D2R Server [2], which allows to specify mappings from relations and attributes to classes and predicates. Due to the variety of attributes we capture, we cannot directly reuse a vocabulary. For instance, *foaf:Person* does not have a middle name. Therefore, we map different entity types to subclasses (we defined) of the following ontology classes:

- person: *foaf:Person*
- legal entity: *foaf:Agent*
- fund: a class specifically defined for our purpose

Due to the integration of Freebase data, the resulting dataset also contains links to Freebase enabling path access to a broad set of further LOD data sources. Thus, GovWild data is well-connected to the LOD cloud, since Freebase is one of its major hubs. The creation of additional links, which can be done with ODDLinker [4], is left for future work.

The generation of GovWild id URIs is based on suggestions by Sheridan and Tennison [8]. Moreover, we incorporated descriptive information. Specifically, we have created URIs of the following form:

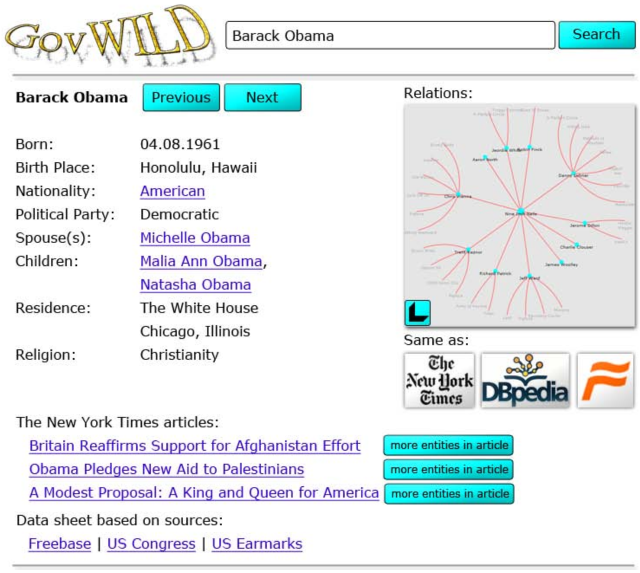
- http://govwild.org/id/person/{first_name}_{middle_name}_{last_name}_{year_of_birth}_{id}
- http://govwild.org/id/legal_entity/{name}_{form}_{id}
- http://govwild.org/id/fund/{subject}_{year}_{id}
- http://govwild.org/def/funding/{concept}

## 5. THE GOVWILD SITE

Data and tool are available at `govwild.hpi-web.de`. The tool as depicted in Fig. 4 is a search engine-like web application that allows to (keyword) query and browse the data interactively. Besides search results we graphically display connected entities from the RDF graph. Additionally, we have connected GovWild data with open NYT data (via Freebase links). With these links at hand, we are able to point to articles that mention one of the entities under consideration. Further, for these NYT articles we present other entities also mentioned but not part of the search result. In addition to the tool and the Linked Data interface (URI lookup) `govwild.hpi-web.de` also offers the dataset as download and allows to run SPARQL queries against it. At the time of writing the data comprises the following:

- 170.600 persons from the US and the EU
- 175.400 legal entities from the US and the EU
- 1.667.200 funds

We have several plans for future work: Currently, we are integrating further sources, namely the German commercial register (describing all German companies) as well as the US Federal Election Commission (comprising campaign finance information), which will greatly enrich GovWild data. Generating additional links, e.g., to GeoNames data, can also shed more insights to GovWild data. Future research will focus on on-the-fly integration while new data is generated.



**Barack Obama**

Born: 04.08.1961
Birth Place: Honolulu, Hawaii
Nationality: American
Political Party: Democratic
Spouse(s): Michelle Obama
Children: Malia Ann Obama, Natasha Obama
Residence: The White House
Chicago, Illinois
Religion: Christianity

The New York Times articles:
Britain Reaffirms Support for Afghanistan Effort
Obama Pledges New Aid to Palestinians
A Modest Proposal: A King and Queen for America

Data sheet based on sources:
Freebase | US Congress | US Earmarks

**Figure 4: The GovWild tool.**

**Acknowledgements.** We thank Howard Ho and Mauricio Hernandez from IBM Almaden Research Center as well as Antonio Sala from University of Modena for their valuable support. Olaf Hartig, Jörg Waitelonis and Harald Sack provided additional feedback.

## 6. REFERENCES

[1] L. Chiticariu, R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, and S. Vaithyanathan. SystemT: An algebraic approach to declarative information extraction. In *48th Annual Meeting of the Association for Computational Linguistics (to appear)*, 2010.

[2] R. Cyganiak and C. Bizer. D2R Server - publishing relational databases on the web as SPARQL endpoints. In *Proc. of 15th Int. World Wide Web Conf.*, 2006.

[3] U. Draisbach. The duplicate detection toolkit (DuDe). `www.hpi.uni-potsdam.de/naumann/projekte/dude`.

[4] O. Hassanzadeh and M. Consens. Linked Movie Data Base. Triplification Challenge, 2008.

[5] A. E. Monge and C. P. Elkan. The field-matching problem: Algorithm and applications. In *Proc. of 2nd Int. Conf. on Knowledge Discovery and Data Mining*, 1996.

[6] A. Sala, C. Lin, and H. Ho. Midas for government: Integration of government spending data on Hadoop. In *Proc. of the Int. Workshop on New Trends in Information Integration (NTII)*, Long Beach, CA, 2010.

[7] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, 1976.

[8] J. Sheridan and J. Tennison. Linking UK government data. In *Proc. of the WWW Workshop on Linked Data on the Web*, 2010.

[9] W. E. Winkler. Overview of record linkage and current research directions. Technical report, Statistical Research Division at U.S. Census Bureau, 2006.