GovWILD: Integrating Open Government Data for Transparency

Christoph Böhm[1], Markus Freitag[2], Arvid Heise[1],
Claudia Lehmann[2], Andrina Mascher[2], Felix Naumann[1]
Hasso Plattner Institute, Potsdam, Germany
[1] firstname.lastname@hpi.uni-potsdam.de
[2] firstname.lastname@student.hpi.uni-potsdam.de

Vuk Ercegovac[1], Mauricio Hernandez[2]
IBM Almaden Research Center, CA, US
[1] vercego@us.ibm.com [2]mauricio@us.ibm.com

Peter Haase, Michael Schmidt
fluid Operations AG, Walldorf, Germany
firstname.lastname@fluidops.com

ABSTRACT

Many government organizations publish a variety of data on the web to enable transparency, foster applications, and to satisfy legal obligations. Data content, format, structure, and quality vary widely, even in cases where the data is published using the wide-spread linked data principles. Yet within this data and their integration lies much value: We demonstrate *GovWILD*, a web-based prototype that integrates and cleanses Open Government Data at a large scale. Apart from the web-based interface that presents a use case of the created dataset at `govwild.org`, we provide all integrated data as a download. This data can be used to answer questions about politicians, companies, and government funding.

1. GOVERNMENT TRANSPARENCY

During the past years, the amount of data provided on the web has increased enormously. However, the quality (correctness, completeness, consistency, etc.) of the data differs widely. Interested individuals want to investigate the provided information, but using and understanding the relevant sources (eParticipation) is a difficult task for a user. In general, there is a high demand for an integrated access to web data and Linked Open Data (LOD). The latter is published in RDF format and follows a set of best practices in order to facilitate easy understanding and reuse [2]. General open web data comes in any format ranging from text or tables in HTML and PDF to CSV files following a sort of schema. Open Government Data rather focuses on openness[1].

Because of the important role of governments, administrations, and the general knowledge they assemble, the value of freely accessible public data from government agencies is especially outstanding. Open Government Data is meant to provide transparency of governmental functions. But also here, the vast amount of government data, schematic heterogeneity, and the lack of consistency complicate access and integration. Therefore, the task of providing unified, structured, and interlinked data is daunting but worthwhile. Published clean data can be analyzed, visualized, or further interconnected. Amongst others, a benefit is a heightened transparency of government actions.

The goal of GovWILD (Government Web Data Integration for Linked Data) is to integrate and interconnect government data by creating links among existent datasets as well as to provide a web-based application that enables exploration of the resulting data. To this end, large amounts of data from the US and the EU are connected with open data from various sources.

Of particular interest in GovWILD is data that is connected with financial data of governments or public funds in general. This includes *persons*, who work for a governmental agency or in any position that is financed with public funds, *companies* and their major employees or shareholders, as well as information about *governmental spending*, which may result in a relationship between the government and certain companies. This data is accessible to the public to easily investigate facts about governments instead of browsing through scattered and unstructured information.

The GovWILD project has been included in the LOD cloud diagram[2] and in CKAN's Data Hub[3], a directory of known LOD sources, with link connections to DBpedia (5,845 entities linked), Freebase (132,953), and the New York Times (6,702).

2. GOVERNMENT DATA IN THE WILD

The exploration of Open Government Data is not an easy task. In a first step, one needs to discover sources that provide information for the task at hand. Then, it requires meaningful connections between datasets. Consider for instance the following information interest: *Which companies profit from cash flows initiated by Barack Obama?*

To tackle this question, one could issue a web search with the keywords *barack obama* and *spending* or *earmarks*. This yields many posts discussing Obama's politics but does not lead to any (semi-)structured information about spendings. Using *earmarks* only, one can find OMB's Earmarks site on `earmarks.omb.gov` where some digging leads to Earmarks issued by Barack Obama and others – see for instance Fig. 1. Here, the recipient is the Department of Defense (DOD), specifically the Rock Island Arsenal. For discov-

[1] `http://www.opengovdata.org/home/8principles`

[2] `http://lod-cloud.net`

[3] `http://thedatahub.org`

World Wide Web, 2012 Lyon, France
ACM 978-1-4503-1230-1/12/04.

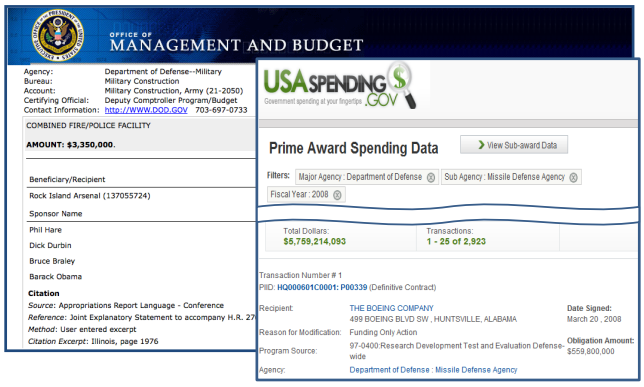

Figure 1: Exemplary searches for open data on government fundings: US Earmarks and USA Spending

ering where the money went from there, one can search `usaspending.gov`. However, the search on this site is very complex and does not allow to specify *Military Construction Bureau*. When searching all military sub-agencies, and thus involving fuzziness, one finds *The Boeing Company* among others. In order to further investigate this company, one could query Freebase, where this company's primary name is simply *Boeing*. Also, when interested in biographic information for Barack Obama, where we started our search endeavor, we could look at `bioguide.congress.gov`, which has textual information about politicians.

Current public open data initiatives support transparency in that they uncover data that has been kept within organizations before. However, these open data sources co-exist as silos scattered across the internet resulting in different formats, structures, and semantics.

With GovWILD, we overcome this heterogeneity and thus facilitate a higher level of transparency by providing a clean and integrated view on the data. So far, we have selected data sources from the US, the EU, and especially Germany that deal with related semantic content and have overlapping time ranges. Further, we have incorporated some general information sites to augment person and legal entity data. GovWILD offers a concise view onto persons, funds, and legal entities as well as the connections between these entities. Table 1 lists the currently integrated datasets.

Data Sources	Years	Records	Attr.	Format
US-Spending	2009	1,724,654	142	XML
US-Earmarks	2008-2009	102,275	38-133	CSV
US-Congress	1774-2009	107,627	8	HTML
EU-Finance	2007-2009	150,499	15	HTML
EU-Parliament	1952-2010	3,133	19	HTML
DE-Party Donations	2000-2009	1,102	5	HTML
DE-Bundestag	2009-2011	629	10	HTML
DE-Agricultural Subsidies	2007-2008	103,652	9	HTML
Freebase	-	1,725,219	94	TSV
New York Times	-	8,013	5	HTML

Table 1: Data sources integrated in GovWILD

The US sources include spendings of federal government agencies to contracted companies (US-Spending), funds for public or private projects enacted by individual congress members (US-Earmarks), and biographical information about congress members including family relationships (US-Congress). Here US-Spending is the largest source we integrated (9 GB in raw XML for one year). Similarly, the EU sources cover spendings (EU-Finance), subsidies (DE-Agricultural Subsidies) and politicians (EU-Parliament, DE-Bundestag). Additionally, we integrated donations from legal entities and private persons to German parties (DE-Party Donations).

Further, we selected chunks from Freebase to augment information about entities from former sources: companies, persons, educations, board memberships, etc. Finally, New York Times articles allow for an investigation of news articles with respect to entities occurring in other sources. In this way GovWILD also incorporates media to uncover more interesting relationships between legal entities and persons. Additionally, the latter two sources connect entities from government data via *owl:sameAs* links to other open general purpose data on the internet. Note that the data GovWILD covers at this point can be considered a sample from the public data currently available. An abstract data model as well as an extensible integration workflow allows for an easy extension with other sources including new types of entities.

A closer inspection of the data reveals integration challenges. On the technical level, most datasets are only available as online web sites and need to be crawled first. Downloadable data sources use different formats to represent the data (CSV, TSV or XML). The quality of data documentations varies largely. Furthermore, the data exhibits a high schematic heterogeneity. All columns need to be mapped to a global schema. Unfortunately, the schemas of US-Earmarks and EU-Finance change quite often and each version has to be mapped individually.

On the semantic level, entities in different data sources need to be identified and fused to complete the integration: Values within and across data sources are denormalized, e.g. the first names of US-Earmarks are often nicknames. Here, missing values and misspellings make the record linkage even more challenging. Finally, conflicts between values of different data sources need to be resolved to form a concise entry.

3. THE INTEGRATION PROCESS

To scale well on the number of data sources, we decided to implement the integration process in several interconnected Jaql scripts that can be run on Hadoop clusters [1]. The integration process consists of source-specific and inter-source data cleansing tasks [6]. Table 2 summarizes the programmatic integration steps.

Preparation: As a prerequisite for programmatic integration, we first select and inspect relevant data sources. Some data sources provide database dumps, which we convert to JSON, the internal data format of GovWILD and Jaql. For web sources that do not provide database dumps, we use specifically configured crawlers with built-in text extraction. Below is an excerpt of the tuple representing Barack Obama in the formatting of the crawled US-Congress dataset:

```
"memberName" : "OBAMA, Barack",
"birthDeath" : "1961- "
```

Scrubbing and mapping: The goal of our source-specific scrubbing scripts is to overcome schematic heterogeneity. Fig. 2 depicts the global schema of GovWILD. The schema is very simple and thus easy to extend if new sources cannot be mapped to existing entities [3].

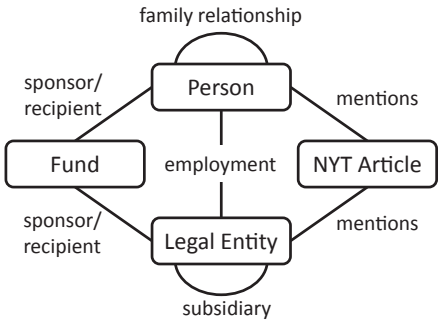

Figure 2: Schema of GovWILD

The scripts perform three cleansing operations. First, invalid values are either corrected or the corresponding tuple is removed. Second, scripts normalize values, e.g., name splitting and conversion of state codes to state names using dictionaries. Third, the scripts extract the entity types and their respective relationships using hand-crafted schema mapping rules, which involve, for example, row grouping in US-Earmarks to collect N:M relationships. The US-Congress tuple about Obama would be mapped to:

```
"firstName" : "Barack",
"lastName" : "Obama",
"birthDate" : { year: 1961 }
```

Entity matching: The next steps identify real-world entities across different data sources. The following two tuples illustrate the goals and challenges. They were extracted from Freebase and represent president Barack Obama and his father.

```
"firstName" : "Barack",
"lastName" : "Obama",
"birthDate" : { day: 4, month: 8, year: 1961 }

"firstName" : "Barack",
"lastName" : "Obama",
"nameAddition" : "Sr",
"birthDate" : { year: 1936 }
```

Despite identical names, these entries obviously represent two different real-world persons. While we would like to find the connection between the first record and Barack Obama from the US-Congress dataset, we need to be careful not to match it with the second entry.

Further, the conceptually quadratic number of comparisons among entities needs to be limited. We use the sorted neighborhood method implemented in the DuDe framework to reduce the number of comparisons and thus the computation time [4]. It sorts all entities with a sorting key, slides a window of size 50 over the data, and compares only those entities that appear in the same window. Multiple passes with different keys increase the probability of indeed comparing matching records.

Step	Time	Size
Jaql 0.5	4h 46min	
Scrubbing and mapping	1h 34min	
Matching of legal entities - Finding similar entities (no MapReduce) - Fusing similar objects - Updating fused IDs	1h 5min 10min 10min	Input: 259K entities Output: 251K entities
Matching of persons - Finding similar entities (no MapReduce) - Fusing similar objects - Removing unconnected persons from Freebase - Updating fused IDs	48min 11min 7min	Input: 1.3M entities Output: 153K entities including 35K fusions
Post-processing - Adding URI and label - Replacing ID references by URI references - Precomputing canned queries	2min 16min 21min	LegalEntity: 251K entities Person: 153K entities Fund: 1M entities
Export JSON to RDF	12h	43M N3 triples
Import into IWB	6h	
Index in IWB	1h 30min	

Table 2: Integration workflow

To decide whether a candidate pair of records represents the same real-world entity, we developed similarity measures that exploit as much source-specific information as possible. US-Earmarks tends to use nicknames, such as *Bill* vs. *William*, as used in many other sources. We especially reject false positives on logical contradictions, e.g., a congress member cannot enact a fund after death. Further constraints help to choose more appropriate matches, e.g., in the example above we choose the younger Obama since birth dates are closer together. The similarity measures for persons use relationships to legal entities, since these were matched and fused before.

Data fusion: The last integration step fuses previously matched and grouped representations of entities to concise tuples in order to lower the complexity of queries spanning multiple entities. The different representations are usually partially complementary and partially overlapping. In the former case, null values are discarded in favor of non-null values. The latter case, however, often results in conflicting values. We employ belief functions from Dempster-Shafer theory to resolve such conflicts. We start with some initial belief in the quality of the data sources and collect evidence for the different values. For instance, *B.* strengthens the belief in the canonical first name *Barack* but not vice versa. In the end, the value with the highest belief is chosen.

Consequently, the resulting entity incorporates references from multiple fused entities. Incoming references need to be updated. Furthermore, the resulting entity contains all lineage information to allow users to trace the origin of the fused entries at hand.

Post-processing: Eventually, the integration script exports all data as RDF triples. Further, the script renders all interconnections found in the integration process as *owl:sameAs* triples to enrich the LOD Cloud.

In parallel, we also prepare the data for the web application. Some canned queries enrich the dataset with complex aggregations (that cannot be expressed in SPARQL 1.0).

Finally, we import the extended data into the Information Workbench (IWB) [5].

4. THE DEMO

GovWILD provides a web frontend to explore the integrated dataset at `govwild.org`. To visualize the relations between persons, funds, and entities from government and industry, GovWILD uses the IWB platform. In addition to the visual interface, GovWILD provides a SPARQL query interface to allow queries such as for the funds that Barack Obama sponsored:

```
SELECT ?recipient ?amount ?description WHERE {
    ?fund ontology:sponsor ?person .
    ?person ontology:name 'Barack Obama' .
    ?fund ontology:recipient ?recipient .
    ?fund ontology:amount ?amount .
    ?fund ontology:description ?description . }
```

In the following, we describe a walkthrough for the above mentioned use case using the explorative interface instead of complex SPARQL queries. The user might start by entering the name *Barack Obama* into the search field and is then presented a list of all corresponding entities from which she might choose the person entity *Barack Obama*. The user is redirected to the wiki view, which we customized for each entity type based on widgets with generic SPAQRL queries. In the wiki view for persons, the user can inspect their biography, *owl:sameAs* links, family and employment relations, and funds. To track data lineage, GovWILD provides links to the original websites and presents the individual entities before fusion. If users wish to filter or sort the data of an entity, they can switch to the table view. The third view visualizes connections between selected entities as a graph with a maximum of two levels, as shown in Fig. 3. Hence, the user can see that Obama sponsored several Earmarks, such as 384,000 USD to the Sparks College. By clicking on *Sparks College* or another node in the graph, the user is redirected to the graph view of the chosen entity.

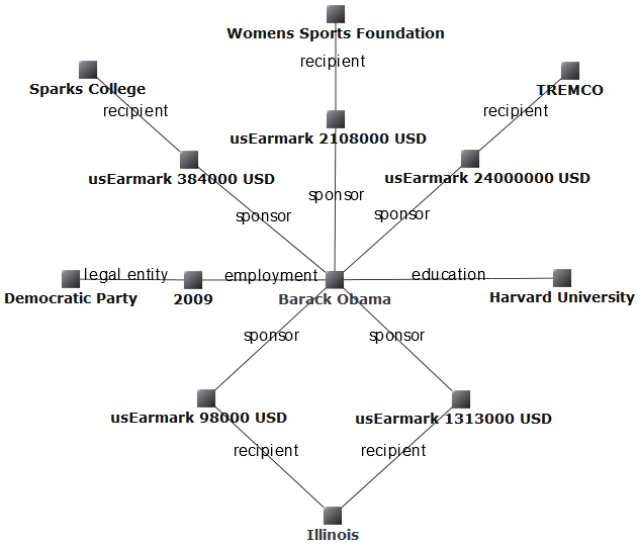

Figure 3: Graph view of Barack Obama with selected connections to funds and legal entities

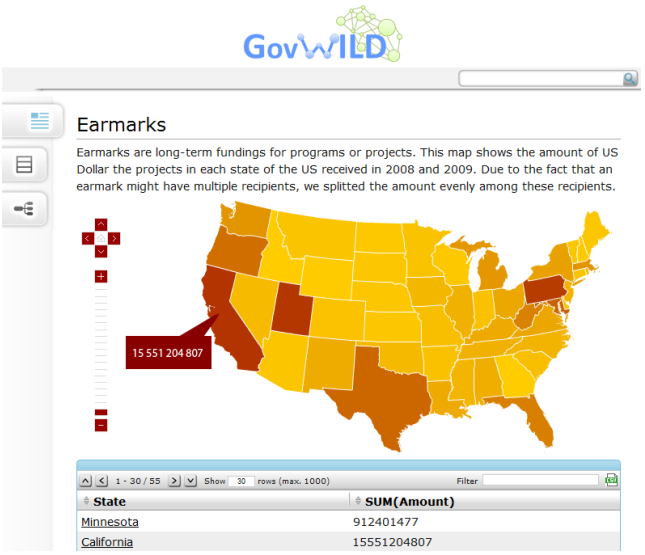

Figure 4: Amount of Earmarks aggregated by US states in a canned query

In addition to exploring only the Earmarks of Barack Obama as in our walkthrough, GovWILD aggregates all Earmarks from 2008 and 2009, as shown in Fig. 4. GovWILD provides such canned queries to demonstrate the potential of our integrated dataset. The data itself can be downloaded as JSON and RDF or retrieved via the SPARQL endpoint for further analysis and to enrich the open data initiative.

Acknowledgements. This research was supported by the German Research Society (DFG grant no. FOR 1306) and an IBM Scalable Data Analytics award.

5. REFERENCES

[1] K. S. Beyer, V. Ercegovac, R. Gemulla, A. Balmin, M. Y. Eltabakh, C.-C. Kanne, F. Özcan, and E. J. Shekita. Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. *PVLDB*, 4:1272–1283, 2011.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – The Story So Far. *Int. Journal on Semantic Web and Information Systems (JSWIS)*, 5:1–22, 2009.

[3] C. Böhm, F. Naumann, M. Freitag, S. George, N. Höfler, M. Köppelmann, C. Lehmann, A. Mascher, and T. Schmidt. Linking open government data: what journalists wish they had known. In *Proc. the Int. Conf. on Semantic Systems, I-SEMANTICS*, 2010.

[4] U. Draisbach and F. Naumann. DuDe: The duplicate detection toolkit. In *Proc. of the Int. WS on Quality in Databases (QDB)*, 2010.

[5] P. Haase, M. Schmidt, and A. Schwarte. The information workbench as a self-service platform for linked data applications. In *Int. WS on Consuming Linked Data (COLD)*, 2011.

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