NIF4OGGD - NLP Interchange Format for Open German Governmental Data

Mohamed Ahmed Sherif[1], **Sandro Coelho**[2], **Ricardo Usbeck**[1,3], **Sebastian Hellmann**[1], **Jens Lehmann**[1], **Martin Brümmer**[1] **& Andreas Both**[3]

[1] Leipzig University, Germany
[2] Federal University of Juiz de Fora, Brazil
[3] R & D, Unister GmbH, Germany

Abstract

In the last couple of years the amount of structured open government data has increased significantly. Already now, citizens are able to leverage the advantages of open data through increased transparency and better opportunities to take part in governmental decision making processes. Our approach increases the interoperability of existing but distributed open governmental datasets by converting them to the RDF-based NLP Interchange Format (NIF). Furthermore, we integrate the converted data into a geodata store and present a user interface for querying this data via a keyword-based search. The language resource generated in this project is publicly available for download and also via a dedicated SPARQL endpoint.

Keywords: NLP, NIF, Open Data

1. Motivation

The open data movement has become increasingly important as a key driver for economical success. For instance, the German government has passed an *E-Government-Law* (Bundesgesetzesblatt, 2013), which emphasises the importance of machine-readable data provided by official agencies and other German government organizations. To use this potential we present our multi-dataset mashup *NIF4OGGD*[1].

Many open data platforms, however, still provide data in (often proprietary, non-standard) formats that lack machine-readability. There is a growing number of data repositories. For example, the city of *Berlin* (see Section 2.) provides more than 200 datasets from kindergarten locations across city districts to ozone pollution distributions. While this diversity allows to create a wide range of apps and mashups, the integration of several sources remains a challenging problem. In order to provide a standardized solution to gather and correlate open data documents, we propose to use formats and tools that achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. A format that fits these requirements is the Linked Data-based Natural Language Processing Interchange Format (NIF) (Hellmann et al., 2012; Hellmann et al., 2013). We employ NIF to connect three different government data repositories and interlink it with spatial information in the Web of Data. The resulting language resource is published along with a user interface for browsing it.

The contributions of this paper are: (1) we describe a process for creating a novel language resource (LR) comprising several open datasets across Germany, which we (2) made publicly available. Furthermore, we (3) use Linked Data via NIF as multi-lingual interchange format to allow queries across datasets. Additionally, we (4) offer a simple search engine interface for end users. Finally, we (5) provide use cases that show the potential impact of NIF4OGGD.

[1] http://aksw.org/Projects/NIF4OGGD

2. State of the Art of Open German Governmental Data

In general, in Germany it is not allowed to publish person-centered data without the consent of the described persons. Taking this restriction into account, most data portals upload statistical or textual data from public hearings. This data is difficult to understand without an extra layer of structural information and most often proprietary, unstructured, not standardized and thus not machine–readable. We aim to overcome these problems via NIF (NLP interchange format) which is based on the principles of *Linked Data*.

Tim-Berners Lee postulated the *5 Star principle* (Berners-Lee, 2011) for sharing open data as Linked Data. To get the first star make the data available on the web via an open licence. For the second star, put it on-line in a structured format. The data will get a third star if it is in a non-proprietary format like CSV and a fourth star if URIs are used to denote resources. Finally, it is 5 Star data if it linked to other datasets creating a richer context.

We analyzed three local portals providing open data with respect to their 5 Star quality. Moreover, we look for the availability of geodata, SPARQL (Prud'hommeaux and Seaborne, 2008) endpoint availability and whether the portal has a visual interface for analysing the data on-site, e.g., a heat map function. The results can be seen in Table 1.

Berlin[2] is the pioneer portal for open data in Germany. Although it does not provide 5 Star data, it comes up with an API and several well structured and non-proprietary data files about, e.g., public wireless LAN locations, events or a list of all memorials. Overall, there are 289 datasets in 21 categories.

Bonn[3] is by now not able to deliver content via download or API. The city of Bonn is discussing how and what data should be delivered. A part of the administrative data is available via an administration management system, cf. Section 3.2..

[2] http://daten.berlin.de/

[3] http://www.bonn.de/rat_verwaltung_buergerdienste/aktuelles/open_data

Portal	⋆	⋆⋆	⋆ ⋆⋆	⋆⋆ ⋆⋆	⋆⋆ ⋆⋆⋆	geodata	SPARQL	visual analytics
Berlin	(✓)	(✓)	(✓)	(✓)	(✓)	✓	✗	✗
Bonn	✓	✗	✗	✗	✗	✗	✗	✗
Cologne	(✓)	(✓)	(✓)	✗	✗	✓	✗	✗
NIF4OGGD	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Different data portals, their 5 Star classification and further features. (✓) means that not all data is available at this particular star level.

Listing 1: Select all streets of Berlin along with latitude and longitude.

```
PREFIX lgd:      <http://linkedgeodata.org/ontology/>
PREFIX geovocab: <http://geovocab.org/geometry#>
PREFIX geo:      <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT  ?s ?streetLabel ?lat ?long
FROM   <http://thedatahub.org/dataset/lgd-berlin>
WHERE{
  ?s         a                    lgd:HighwayThing;
             rdfs:label           ?streetLabel;
             geovocab:geometry    ?geometry.
  ?geometry  lgd:posSeq           ?posSeq.
  ?posSeq    ?posSeqP             ?posSeqO.
  ?s2        geovocab:geometry    ?posSeqO;
             geo:lat              ?lat;
             geo:long             ?long.
}
```

Cologne[4] offers 172 datasets from 9 categories for download. Although, the portal provides several open license datasets, no dataset exists following the Linked Data paradigm. Like in the case of *Bonn*, administrative data is available via an administration management system that already has been scraped by the Cologne Open Data Portal[5].

3. Dataset

In this section, we briefly present *LinkedGeoData* (Stadler et al., 2012) and our data extraction from E-Government data portals. Moreover, we point out current problems and how we overcome each of them by using NIF and the Linked Data paradigm.

3.1. LinkedGeoData

The *OpenStreetMap* (OSM)(Haklay and Weber, 2008) project offers a freely available and rich source of spatial data. OSM consists of more than 1 billion nodes and 100 million ways stored in a relational database. *LinkedGeoData* (LGD)[6] provides a transformation of OSM data into RDF (Auer et al., 2009), which comprises approximately 20 billion triples. LGD is available according to the Linked Data principles and interlinked with *DBpedia* (Auer et al., 2008) and *GeoNames*[7]. LGD provides its RDF data not only in form of free dump files , but the data can also be queried via a SPARQL endpoint[8]. LGD provides an ontology for structuring the information in OpenStreetMap. For instance, it contains more than fourty subclasses of `HighWay`. As an example, Listing 1 is a SPARQL query, which retrieves all streets of the city of *Berlin*, along with latitude and longitude information.

For obtaining the relevant datasets, we downloaded OSM dumps[9] for *Berlin* and *North Rhine-Westphalia* and applied the LGD conversion[10] to them. The resulted datasets are stored in the project endpoint[11].

3.2. Data Extraction

To enrich the spatial data, government data was retrieved from administration management systems of Bonn and Cologne, see Section 2.. All of the data is document based, containing *Portable Document Format* (PDF) documents with administrative decisions and documented enquiries of citizens. Each document serves as a resource in the administration management systems and features a title and minor metadata. To obtain the data, the portals were queried by custom web scrapers.

In the case of *Bonn*, the PDF documents where downloaded[12] via `curl` (Steinberg, 2012) and converted to textual data via Linux command `pdf2text`. Because of this procedure, the original formatting of the documents and any information contained in it (like tabular data) was lost. The negative effect of this is limited in our case, since we

[4] `http://www.offenedaten-koeln.de/`
[5] `http://offeneskoeln.de/`
[6] `http://linkedgeodata.org/`
[7] `http://www.geonames.org/ontology`
[8] `http://linkedgeodata.org/sparql`
[9] `http://geofabrik.de`
[10] `https://github.com/GeoKnow/LinkedGeoData`
[11] `http://mlode.nlp2rdf.org/sparql`
[12] `http://www2.bonn.de/bo_ris/ris_sql/agm_index.asp`

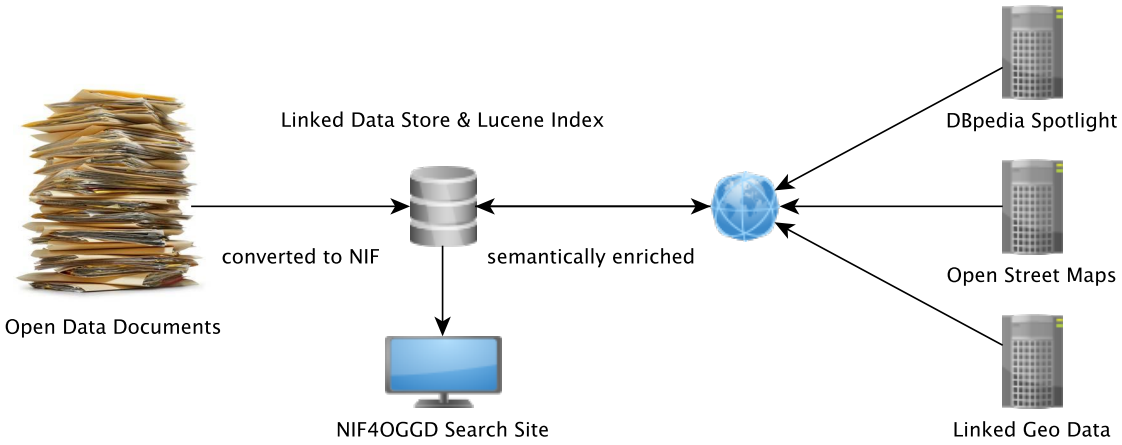

Figure 1: Architecture of the NIF4OGGD system

only perform text search over the documents. However, keeping this information and performing advanced extraction methods is one of our steps in a larger research agenda. In case of *Cologne*, we applied the above mentioned approach which allow us to extract textual data from the documents as well as documents' titles.

4. Architecture

The NIF4OGGD architecture shown in Figure 1 has three main modules: Conversion of documents to NIF, Enrichment and Visualization & Search. The modules provide a flexible solution to integrate multiple web data sources using semantic web standards and NIF.

4.1. Conversion of Documents to NIF

For the conversion to NIF, two resources where established for every document: First, the textual content of a document was added to a resource of the type `nif:Context` with the `nif:isString` predicate as a literal. The URL of the source document was also included in the metadata to ensure full traceability of the original data. Second, the title string of the document was in turn added to a resource of the type `nif:Title` with the `nif:anchorOf` predicate and linked to the context resource via `nif:referenceContext`. Listing 2 provides an example of a NIF conversion of a document presented in Figure 3.

4.2. Enrichment

In order to link the geographical and the governmental data, the governmental data is enriched by geodata via LingPipe Exact Dictionary-Based Chunker (i. LingPipe, 2013). LingPipe relies on the Aho-Corasick string matching algorithm (Aho and Corasick, 1975). The output of this extracting is stored as standardized NIF files. The collection of LGD locations and government data is additionally stored as a set of documents in a *Lucene* (Hatcher et al., 2004) index.

Furthermore, we built an in-memory dictionary for the data provided by LGD. This dictionary is used for performing an analysis in all NIF files that contain governmental data. Specifically in the data stored in `nif:isString` property, when a location name occurs in a document, all the data are stored in the *Lucene* document corresponding to that location. Figure 2 shows an example of all fields stored for an indexed document.

4.3. Visualization & Search

Aiming to allow an easy integration of NIF4OGGD into external web processes, we implemented RESTful and SOAP web services for the search process. The web service interface allows access to query a Lucene index and returns JSON and NIF format.

NIF4OGGD is deployed as a web service and has a very simple user interface for demonstration. The source code is available at our project repository[13]. The user interface[14] was built using the Google Maps API, allowing to search locations by name. Once the location is selected by the user, the map shows it on a map and displays related government documents stored in the index. It is also possible to annotate the texts using DBpedia Spotlight (Mendes et al., 2011) and use the annotated resources to discover more information in a graph database. Figure 3 shows the NIF4OGGD user interface.

5. Use Cases

In this section, we outline application scenarios and use cases for our system.

Data Retrieval. An example use case for data retrieval are citizens searching for events in their neighbourhood. NIF4OGGD provides data about what governmental events happen in a specific area. For instance, Figure 3 shows a user query for governmental documents mentioning *Neumarkt* in *Cologne*. As shown, there are many documents about the search topic, in which the user is free to browse.

Interoperability using NIF. The aligned governmental documents using NIF representation enable searching for the same entity across different authorities' documents. *NIF*'s interoperability capabilities make it easy to query all

[13] `https://github.com/aksw/nif4oggd`
[14] `http://nif4oggd.aksw.org`

Field	IdfpoPSVBNtxx#txxDtxx	Norm	Value
city	Idfp--S--Nnum--------	1.0	Koeln
description	Idfp--S--Nnum--------	1.0	Neumarkt
document	Idf---SV-Nnum--------	0.005E	Antrag auf Aufstellung eines Bebauungsplanes in KölnFlittard, Pützl
document	Idf---SV-Nnum--------	0.005E	Eigenbetriebsähnliche Einrichtung Veranstaltungszentrum Köln Wirt
document	Idf---SV-Nnum--------	0.005E	Baubeschluss zur Realisierung der Pilotanwendung einer umweltser
document	Idf---SV-Nnum--------	0.005E	Ergebnis der Lärmmessungen am Brüsseler Platz@deStellungnahm
document	Idf---SV-Nnum--------	0.005E	Aufwertung der Veedel im Stadtbezirk Innenstadt zu Stadtteilen@de
latitude	Id----S--------------	---	50.9358#50.9359#50.9358#50.9358#50.9358#50.9358#50.9358
longitude	Id----S--------------	---	6.94708#6.94709#6.94799#6.94839#6.94839#6.94633#6.94896
types	Idfp--S--Nnum--------	1.0	DBpedia:Street
url	Id----S--------------	---	http://linkedgeodata.org/triplify/way10807067

Figure 2: Lucene index

Listing 2: Example NIF resources

```
<http://offeneskoeln.de/dokumente/3819-2011/#char=0,1596>
    a             nif:Context , nif:RFC5147String ;
    nif:isString "Baubeschluss zur Realisierung der Pilotanwendung einer umwelt[...]"@de ;
    nif:sourceUrl <http://offeneskoeln.de/dokumente/3819-2011/> ;
    nif:sourceUrl <http://ratsinformation.stadt-koeln.de/vo0050.asp?__kvonr=29978> .

<http://offeneskoeln.de/dokumente/3819-2011/#char=0,40>
    a                    nif:Title , nif:RFC5147String ;
    nif:beginIndex       "0"^^xsd:nonNegativeInteger ;
    nif:endIndex         "12"^^xsd:nonNegativeInteger ;
    nif:referenceContext <http://offeneskoeln.de/dokumente/3819-2011/#char=0,1596> ;
    nif:anchorOf         "Baubeschluss"@de .
```

occurrences of a certain text segment in all available documents without the need of any additional indexing. Especially, sophisticated SPARQL queries can posed to interlinked NIF documents, e.g., to find all places across Berlin where demonstrations were declared about some particular political issue.

Listing 3 introduces an example of using NIF to retrieve all documents mentioning *Baubeschluss*.

Information Aggregation. Using *DBpedia Spotlight*, NIF4OGGD annotates the governmental documents using the cross-domain dataset of *DBpedia*, which provides added value to the data. As an example, in Figure 3 user can use the annotated entities of the presented document to get detailed data about *Kölner Innenstadt*.

6. Conclusion and Future Work

NIF4OGGD is a novel language resource providing geographical data that is interlinked with governmental information. We presented an extraction process for creating this language resource and made it freely available. In addition, our project constitutes a new central sharing point for Open German Governmental Data which is published following the 5 Star principles. In the future, we plan to extend our datasets and enrich it with more data from the LOD-cloud. Furthermore, we plan to employ additional NLP-algorithms, e.g., to just show documents with a valid time range, to provide more trustful documents via NIF4OGGD.

7. References

Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: An aid to bibliographic search. *Commun. ACM*, 18(6):333–340, June.

Sören Auer, Chris Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2008. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer.

Sören Auer, Jens Lehmann, and Sebastian Hellmann. 2009. LinkedGeoData - adding a spatial dimension to the web of data. In *Proc. of 8th International Semantic Web Conference (ISWC)*.

Tim Berners-Lee. 2011. Linked data-design issues (2006). *URL http://www.w3.org/DesignIssues/LinkedData.html*.

Bundesgesetzesblatt. 2013. E-government gesetz des bundes (2013). *URL http://www.bmi.bund.de/SharedDocs/Downloads/DE/Themen/OED_Verwaltung/Informationsgesellschaft/egovg_verkuendung.pdf*.

Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18.

Erik Hatcher, Otis Gospodnetic, and Michael McCandless. 2004. Lucene in action.

Sebastian Hellmann, Jens Lehmann, and Sören Auer. 2012. Linked-data aware uri schemes for referencing text frag-

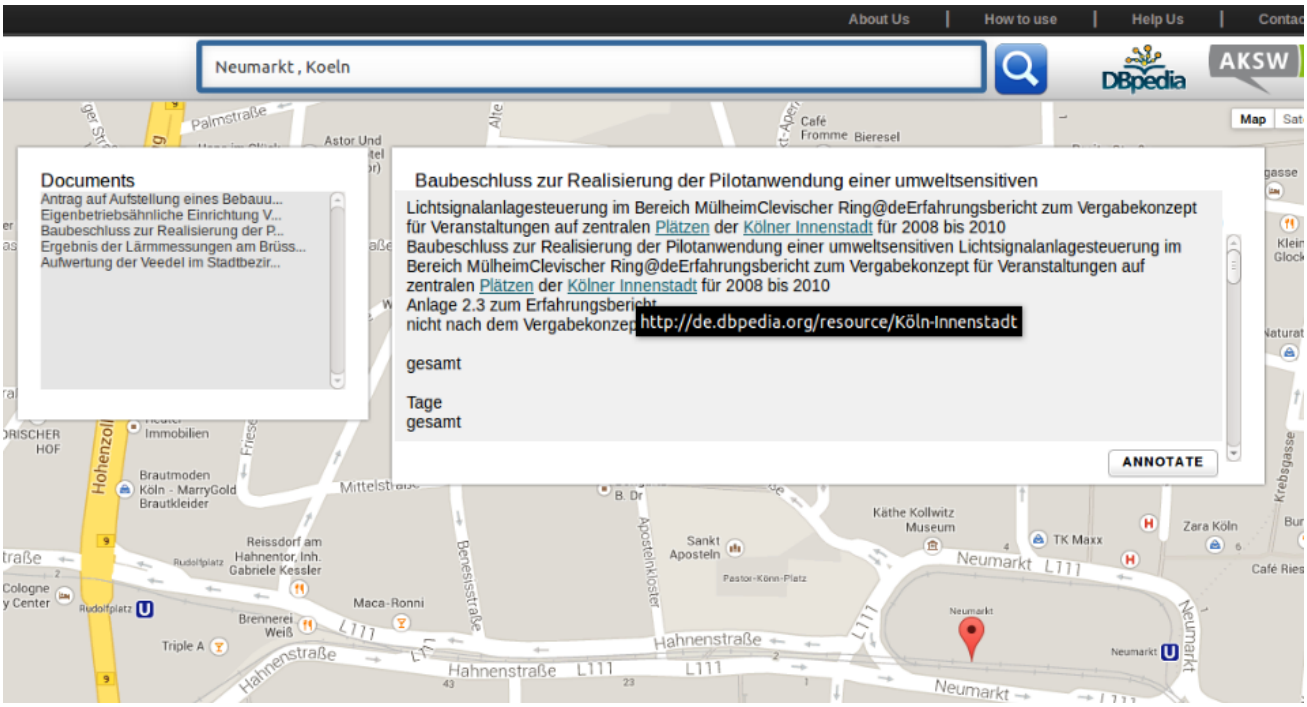

Figure 3: Searching for governmental documents mentioning Neumarkt in Cologne

Listing 3: List of all occurrences of “Baubeschluss” using NIF

```
PREFIX str: <http://nlp2rdf.lod2.eu/schema/string/>
SELECT ?document  {
  ?s            str:isString          ?document.
  ?textSegment  str:referenceContext  ?s;
                str:anchorOf          "Baubeschluss".
}
```

ments. In *EKAW 2012*, Lecture Notes in Computer Science (LNCS) 7603. Springer.

Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. Integrating nlp using linked data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*.

Alias i. LingPipe. 2013. Alias-i lingpipe. [Online; accessed 23-October-2013;].

Pablo N Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM.

Eric Prud’hommeaux and Andy Seaborne. 2008. Sparql query language for rdf. Latest version available as `http://www.w3.org/TR/rdf-sparql-query/`, January.

Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. 2012. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354.

Daniel Steinberg. 2012. libcurl: The multiprotocol file transfer library.

Acknowledgments

Parts of this work were supported by the ESF, the Free State of Saxony and the EU project GeoKnow (GA no. 318159).