The German DBpedia: A Sense Repository for Linking Entities

Sebastian Hellmann, Claus Stadler, and Jens Lehmann

Abstract The modeling of lexico-semantic resources by means of ontologies is an established practice. Similarly, general-purpose knowledge bases are available, e.g. DBpedia, the nucleus for the Web of Data. In this section, we provide a brief introduction to DBpedia and describe recent internationalization efforts (including the creation of a German version) around it. With DBpedia serving as an entity repository it is possible to link the Web of Documents with the Web of Data via DBpedia identifiers. This function is provided by DBpedia Spotlight, which we briefly introduce. We then show how the NLP Interchange Format (NIF) can be used to represent this linking transparently for applications. Using the OLiA ontologies to represent linguistic annotations, NIF allows to represent the output of NLP tools, such as DBpedia Spotlight, in a uniform way.

1 DBpedia

DBpedia (Lehmann et al, 2009; Auer et al, 2008) is a community effort to extract structured information from Wikipedia and to make this information available on the Web. The main output of the DBpedia project is a data pool that (1) is widely used in academics as well as industrial environments, that (2) is curated by the community of Wikipedia and DBpedia editors, and that (3) has become a major crystallization point and a vital infrastructure for the Web of Data. DBpedia is one of the most prominent Linked Data examples and presently the largest hub in the Web of Linked Data (Fig. 1). The extracted RDF knowledge from the English Wikipedia is published and interlinked according to the Linked Data principles.

Sebastian Hellmann · Claus Stadler · Jens Lehmann
Universität Leipzig, Fakultät für Mathematik und Informatik, Abt. Betriebliche Informationssysteme, Johannisgasse 26, 04103 Leipzig, Germany e-mail: hellmann@informatik.uni-leipzig.de

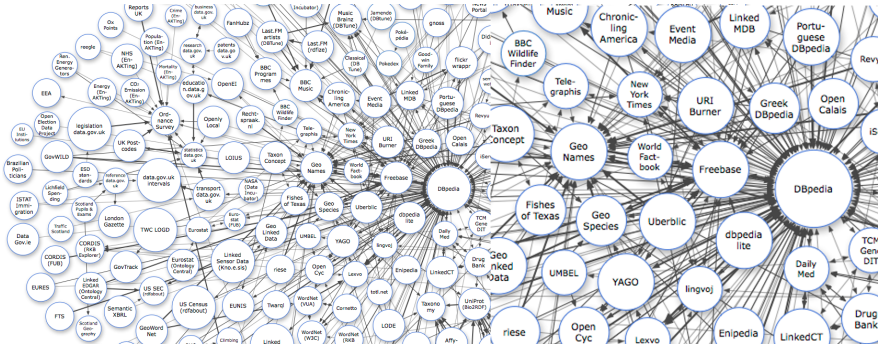

Fig. 1 Excerpt of the data sets interlinked with DBpedia. Source: http://lod-cloud.net (with kind permission of Anja Jentzsch and Richard Cyganiak)

According to Bizer (2011), the latest version of DBpedia (3.7 at the time of writing) contains more than 3.64 million things. Among them are 416,000 persons, 526,000 places, 106,000 music albums, 60,000 films, 17,500 video games, 169,000 organizations, 183,000 species and 5,400 diseases, which are classified in a consistent ontology. Multilingual labels and abstracts in about 97 different languages are provided. Furthermore, there are 2,724,000 links to images, 6,300,000 links to external web pages, 6,200,000 external links into other RDF datasets, and 740,000 Wikipedia categories. Approximately 385 million RDF triples were extracted from the English Wikipedia, whereas an additional 665 million ones originate from other language editions and links to external datasets. As such, the dataset amounts to roughly 1 billion triples.

Currently, the DBpedia Ontology is maintained in a crowd-sourcing approach and thus freely editable on a *Mappings Wiki*:[1] There contributors can 1) edit mapping rules about how Wikipedia infoboxes relate to OWL classes and properties, and 2) edit the ontology, such as adding class labels or modelling subsumption hierarchies. An example is shown in Fig. 2: For the article "Mount Everest", a corresponding DBpedia resource is created, which is classified as `dbp-owl:Mountain`. This is because the article contains the infobox "Infobox_Mountain", for which a contributor to the Mappings Wiki stated this rule. The key-value pairs used in the infobox are mapped to RDF triples via further mapping rules.

2 Internationalization of DBpedia

Early versions of the DBpedia Information Extraction Framework (DIEF) only used the English Wikipedia as their sole source. Yet, since then, its focus has shifted to build a fused dataset by integrating information from many different Wikipedia editions. The emphasis of this fused DBpedia is still on the English Wikipedia as it is

[1] `http://mappings.dbpedia.org`

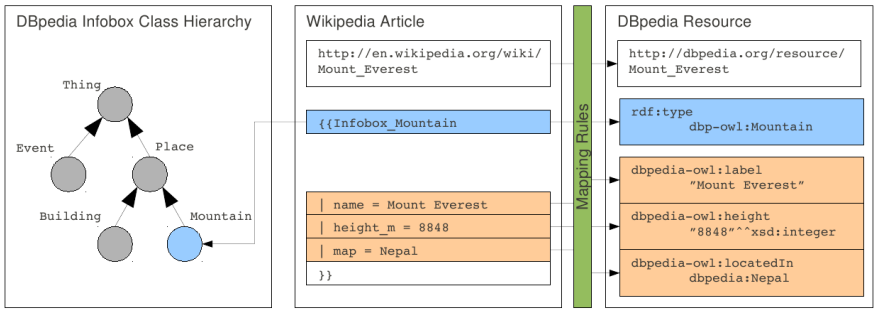

Fig. 2 Rule-based manipulation of extracted data in DBpedia Mappings Wiki

the most abundant language edition. During the fusion process, however, language-specific information was lost or ignored. The aim of the current research in internationalization (Kontokostas et al, 2011) is to establish best practices (complemented by software) that allow the DBpedia community to easily generate, maintain and properly interlink language-specific DBpedia editions. In a first step, we realized a language-specific DBpedia version using the Greek Wikipedia as a basis for prototypical development (Kontokostas et al, 2011). Soon, however, the approach was generalized and applied to 15 other Wikipedia language editions (Bizer, 2011), amongst them the localized German DBpedia. The German Wikipedia is currently the second largest Wikipedia[2] with about 1.3 million articles. With the current version of DIEF, it is responsible for the third largest localized DBpedia with a total of 73 million RDF triples (German being the second largest one). The German community at the DBpedia Mappings Wiki has started to create mappings for infoboxes and achieved a coverage of about 52.89%.[3] A summary of the number of triples extracted for the German DBpedia, hosted by the Freie Universität Berlin[4], is shown in Tab. 1.

Most importantly, DBpedia provides background knowledge for around 3.64 million entities (1.3 million in German) with a high stability with respect to the identifier-to-sense assignment (Hepp et al, 2007). This means that once a piece of text is correctly linked to DBpedia identifier representing the sense, it can be expected that this assignment remains stable.

[2] accessed on Dec 7th, 2011, `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

[3] accessed on Dec 8th, 2011, `http://mappings.dbpedia.org/server/statistics/de/`

[4] `http://de.dbpedia.org`

Filename	Triples
page_links_de.nt	41.395.828
infobox_properties_de.nt	11.055.387
wikipedia_links_de.nt	6.539.427
persondata_de.nt	2.996.640
images_de.nt	2.329.617
labels_de.nt	2.180.233
mappingbased_properties_de.nt	1.764.684
long_abstracts_de.nt	1.137.481
short_abstracts_de.nt	1.137.481
instance_types_de.nt	804.913
interlanguage_links_de.nt	694.064
disambiguations_de.nt	636.481
pnd_de.nt	152.831
homepages_de.nt	117.856
specific_mappingbased_properties_de.nt	103.763
infobox_property_definitions_de.nt	21.510
geo_coordinates_de.nt	36
total	73.068.232

Table 1 Statistics of the extracted triples for the German DBpedia.

3 DBpedia Spotlight

One example application of DBpedia is DBpedia Spotlight (Mendes et al, 2011), a tool for annotating mentions of DBpedia resources in natural language, providing a solution for linking text to the Linked Open Data cloud through DBpedia. DBpedia Spotlight performs term extraction, maps terms to candidate resources and automatically selects a resource based on the context of the input text.

The most basic term extraction strategy in DBpedia Spotlight is based on a dictionary of known DBpedia resource names extracted from page titles, redirects and disambiguation pages. These names are shared in the DBpedia Lexicalization Dataset.[5] The graph of labels, redirects and disambiguations in DBpedia is used to extract a lexicon that associates multiple surface forms to a resource and interconnects multiple resources to an ambiguous name.

First, Wikipedia page titles can be seen as community-approved names. Further, redirects to URIs indicate synonyms or alternative surface forms, including common misspellings and acronyms. And finally, disambiguations provide ambiguous names that are "confusable" with all resources they link to. Their labels (after basic cleaning of words within parenthesis) become names for all target resources in the disambiguation page. In order to score the association between names and DBpedia resources, page links in Wikipedia are used. For each page link, one association between a name in the anchor text with the resource in the target page is counted. Based on this statistics, a number of scores have been derived and shared in the DBpedia Lexicalization dataset.

[5] http://wiki.dbpedia.org/Lexicalizations

Other term extraction techniques already available through DBpedia Spotlight include Keyphrase Extraction (Frank et al, 1999), a non-lexicalized segmentation approach using a shallow parser, and Named Entity Recognition of People, Locations and Organizations based on OpenNLP.

The disambiguation strategy used in DBpedia Spotlight 0.5 models each DBpedia resource in a vector space model of words extracted from Wikipedia paragraphs containing page links. A paragraph is added to the context of a DBpedia resource if that resource's corresponding Wikipedia page is the target of a page link in that paragraph. The words in the paragraph are further scored based on the TF*ICF measure (Mendes et al, 2011), which scores words on their ability to distinguish between known senses of a given term.

DBpedia Spotlight allows users to configure the annotation behaviour based on a number of parameters. Through the use of the DBpedia Ontology, the system allows users to restrict the annotations to a set of classes (e.g. Politician, Restaurants, etc.), or to any arbitrary query expressed in SPARQL. Furthermore, a number of scores, including prominence, contextual pertinence and confidence allows users to deal with the natural trade-off between precision and recall (Gordon and Kochen, 1989).

One recent development is the internationalization of DBpedia Spotlight and the development of entity disambiguation services for German and Korean has begun. Other languages will follow soon including the evaluation of the performance of the algorithms in other languages.

4 NLP Interchange Format

The NLP Interchange Format (NIF)[6] brings together the previously described concepts. It is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations. The core of NIF consists of a vocabulary, which can represent strings as RDF resources. A special URI design is used to pinpoint annotations to a part of a document. These URIs can then be used to attach arbitrary annotations to the respective character sequence. Based on these URIs, annotations can be interchanged between different NLP tools and applications. NIF consists of 3 components:

1. Structural Interoperability (cf. Sect. 4.1): URI recipes are used to anchor annotations in documents with the help of fragment identifiers. The URI recipes are complemented by two ontologies (String Ontology and Structured Sentence Ontology), which are used to describe the basic types of these URIs (String, Document, Word, Sentence) as well as the relations between them (sub/super string, next/previous word).
2. Conceptual Interoperability: Best practices for annotating these URIs are given to provide interoperability. OLiA,[7] as presented in Chiarcos (this vol.), is used

[6] Specification 1.0: `http://nlp2rdf.org/nif-1.0`

[7] `http://purl.org/olia`

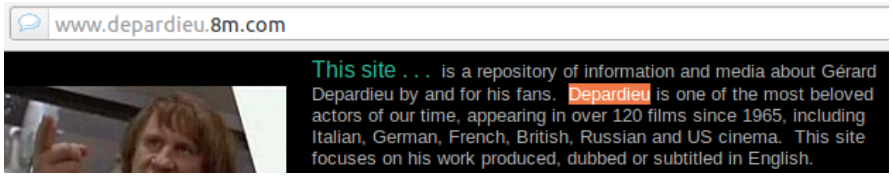

Fig. 3 The second occurence of `Depardieu` is highlighted and is linked in the example with the French DBpedia resource about Gérard Depardieu. Source: http://www.depardieu.8m.com/

for the grammatical features, while the SCMS Vocabulary[8] and DBpedia are used for sense tagging.
3. Access Interoperability: An interface description for NIF Components and Web Services allows NLP tools to interact on a programmatic level.

4.1 Anchoring Web Annotations

One basic use case of NIF is to allow NLP tools to exchange annotations about (Web) documents in RDF. The first prerequisite for achieving this is that strings in documents can be referred to by URIs, so they can be used as a subject in RDF triples. A quite simple example is depicted in Fig. 3, which can be addressed with two possible URI recipes according to the NIF 1.0 specification:

1. A URI scheme with offsets can be used, which is in general easy to compute and handle programmatically:
`http://www.depardieu.8m.com/#offset_22295_22304_Depardieu`
Note that the HTML document is treated as a string or a character sequence. The # is used in this case to address a fragment of the whole document, hence the naming "fragment identifier". 22295 denotes the offset of the first character in the (source code of) the website, whereas 22304 is the last character. "Depardieu" is an addition to make the URI more readable.
2. A URI scheme based on the context and md5 hashes, which is more stable compared to the previous recipe, could be used:
`http://www.depardieu.8m.com/#hash_6_9_e7146a74239c3878aedf0c45c6276618_Depardieu` Considering the example "`nbsp; (Depardieu) is on`", here the context is 6 characters before and after the occurrence of the substring in question. A hash from this context can be computed and added to the URI. The larger the context, the more likely it is that two different strings are not assigned the same URI. The advantage of hashes is that URIs may stay valid even when the document changes.

A NIF 1.0 model, which links the second occurrence of *Depardieu* to the French DBpedia could contain the following RDF triples:

[8] `http://scms.eu`

```
@prefix :       <http://www.depardieu.8m.com/#>
@prefix fr:     <http://fr.dbpedia.org/resource>
@prefix str:    <http://nlp2rdf.lod2.eu/schema/string/> .
@prefix scms:   <http://ns.aksw.org/scms/> .

:offset_22295_22304_Depardieu
    scms:means  fr:Gerard_Depardieu ;
    rdf:type    str:OffsetBasedString .

:offset_0_54093_%3C!doctype%20html%20publi
    rdf:type    str:OffsetBasedString ;
    rdf:type    str:Document ;
    str:subString :offset_22295_22304_Depardieu;
    str:sourceUrl <http://www.depardieu.8m.com/> .
```

Similarly, the output of NLP tools can be represented, e.g., by associating *Depardieu* with its language (e.g., a Glottolog or lexvo identifier), with a syntactic parse tree (e.g., as specified in POWLA), or with morphosyntactic annotations (as provided by OLiA).

4.2 Integration of NLP Tools Using NIF

Many existing NLP tools can be integrated with NIF by providing a wrapper that serializes and deserializes between NIF and its native format. Our experience is, that the implementation efforts of such wrappers is limited, thereby resulting in low integration costs. An NLP pipeline can then be formed by either exchanging NIF models directly, or by merging the NIF models generated by tools that take documents as input.

The NIF wrappers and the special URI recipes ensure that equal strings in equal contexts are assigned equal URIs. Therefore, the integration of the outputs of NLP tools providing annotations for the same part of a document can be done by simply merging the NIF models.

NIF Webservices

We have implemented NIF wrappers for Stanford Core NLP,[9] Apache OpenNLP.[10] Snowball Stemmer,[11] DBpedia Spotlight, and Monty Lingua.[12] In order to minimize the overhead of their reuse, we additionally expose the functionality as RESTful web

[9] `http://nlp.stanford.edu/software/corenlp.shtml`

[10] `http://incubator.apache.org/opennlp`

[11] `http://snowball.tartarus.org/`

[12] `http://web.media.mit.edu/~hugo/montylingua/`

services. Detailed information and documentation about these services can be found at the NLP2RDF website.[13] Online demos are also provided.[14]

Figure 4 shows an example about the merged output of the Snowball Stemmer together with Stanford Core for the sentence "My favourite actor is Natalie Portman!". The horizontally aligned boxes at the center of the image denote the URIs referencing the individual substrings of the source sentence, whereas the attached annotations originate from the Snowball stemmer (arrows labelled `Stem`), and the Stanford framework (`Pos Tag, Type, Lemma`). Additionally, the model is enriched with NIF metadata, such as the successor relation between the constituent words of the sentence (not shown in this figure).

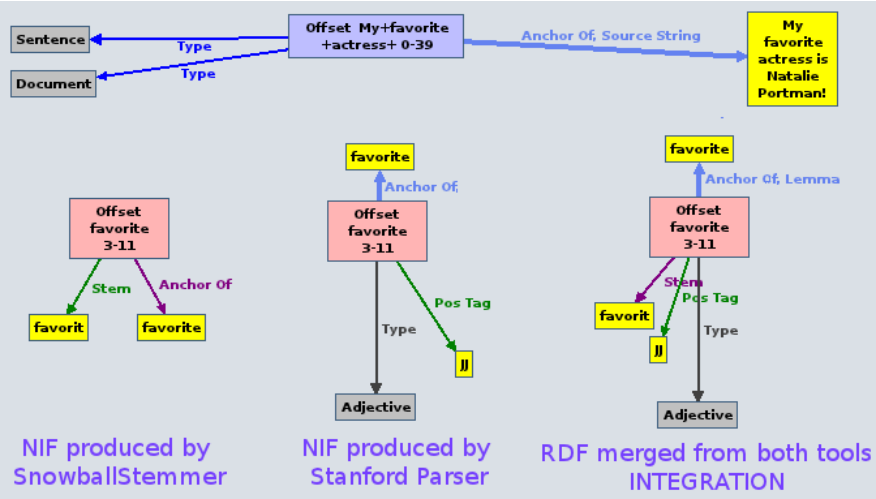

Fig. 4 Merged output of Snowball Stemmer and Stanford.

5 Summary

We have presented several technologies, which together show how text can be annotated using background knowledge from the Linked Open Data cloud:

- DBpedia as cross domain knowledge base.
- Different language editions of DBpedia, in particular the German DBpedia.
- DBpedia Spotlight, as a tool for annotating mentions of DBpedia entities in text.
- NIF as format for interchanging annotations.

[13] `http://nlp2rdf.org`
[14] `http://nlp2rdf.lod2.eu/demo.php`

Overall, we argue that the (German) DBpedia can become a sense repository for annotating entities. In future work, we will also look at the integration of special entities through knowledge bases like LinkedGeoData (Auer et al, 2009; Stadler et al, 2011). NIF provides an easy way to reuse resources from the LOD cloud in general and allows a seamless integration of several NLP tools.

Acknowledgements We would like the thank Pablo Mendes for his contributions to this work.

References

Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z (2008) DBpedia: A nucleus for a web of open data. In: Proceedings of the 6th International Semantic Web Conference (ISWC), Springer, Lecture Notes in Computer Science, vol 4825, pp 722–735, DOI doi:10.1007/978-3-540-76298-0_52

Auer S, Lehmann J, Hellmann S (2009) LinkedGeoData - adding a spatial dimension to the web of data. In: Proc. of 8th International Semantic Web Conference (ISWC), DOI doi:10.1007/978-3-642-04930-9_46, URL http://www.informatik.uni-leipzig.de/~auer/publication/linkedgeodata.pdf

Bizer C (2011) Dbpedia 3.7 released, including 15 localized editions. http://blog.dbpedia.org/2011/09/11/dbpedia-37-released-including-15-localized-editions/

Chiarcos C (this vol.) Interoperability of corpora and annotations. P. 161-179

Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG (1999) Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, IJCAI '99, pp 668–673, URL http://dl.acm.org/citation.cfm?id=646307.687591

Gordon M, Kochen M (1989) Recall-precision trade-off: A derivation. Journal of the American Society for Information Science 40(3):145–151, URL http://www3.interscience.wiley.com/cgi-bin/jtoc/27981/

Hepp M, Siorpaes K, Bachlechner D (2007) Harvesting wiki consensus: Using wikipedia entries as vocabulary for knowledge management. IEEE Internet Computing 11(5):54–65, URL http://dblp.uni-trier.de/rec/bibtex/journals/internet/HeppSB07

Kontokostas D, Bratsas C, Auer S, Hellmann S, Antoniou I, Metakides G (2011) Towards linked data internationalization - realizing the greek dbpedia. In: Proceedings of the ACM WebSci'11

Lehmann J, Bizer C, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) DBpedia - a crystallization point for the web of data. Journal of Web Semantics 7(3):154–165, DOI doi:10.1016/j.websem.2009.07.002, URL http://jens-lehmann.org/files/2009/dbpedia_jws.pdf

Mendes PN, Jakob M, García-Silva A, Bizer C (2011) Dbpedia spotlight: Shedding light on the web of documents. In: Proc. 7th International Conference on Semantic Systems (I-Semantics)

Stadler C, Lehmann J, Höffner K, Auer S (2011) Linkedgeodata: A core for a web of spatial open data. Semantic Web Journal