# A *lemon* lexicon for DBpedia

Christina Unger, John McCrae, Sebastian Walter, Sara Winter, and Philipp Cimiano

Semantic Computing Group
CITEC, Bielefeld University, Germany

**Abstract.** As the body of knowledge available as linked data grows, so does the need to provide methods that make this knowledge accessible for humans. Such methods usually require knowledge about how the vocabulary elements used in the available ontologies and datasets are verbalized in natural language. This has lead to much interest in the development of models and frameworks for publishing ontology lexica as linked data. In this paper we describe a process for the manual development of such lexica in *lemon* format and illustrate some of the key challenges involved. As a proof of concept, we provide a manually created English lexicon for the DBpedia ontology and describe its first release.

**Keywords:** Ontology lexicalization, DBpedia

## 1 Introduction

As the body of knowledge available as linked data grows, so does the need to provide methods that make this knowledge accessible for humans, for example by systems that transform natural language questions into SPARQL queries, systems that generate natural language paraphrases of SPARQL queries, or systems that generate verbalizations of a given ontology or natural language summaries from RDF datasets. All such systems require knowledge about how the vocabulary elements used in the available ontologies and datasets are verbalized in natural language, in particular covering different verbalization variants, possibly in multiple languages. Although standards such as RDFS and SKOS, or models such as OTR [8], allow terminological and linguistic information to be attached to an ontology, this information is very limited and often not rich enough for natural language applications. For example, the DBpedia property `team` can be verbalized as *play for* in English if the subject is any kind of player, while it would be verbalized as *race for* in case the subject is a race driver, and as *manage* in case the subject is a sports manager. Moreover, not only verbalizations of single classes or properties are relevant, but also verbalizations of complex constructions. For instance, the expression *grandchildren* refers to the property chain `child` ∘ `child`. Similarly, the adjective *female* describes individuals that are related to the resource `Female` through the property `gender`. There are many more examples, because commonly the conceptual granularity of natural language and that of the schema underlying a particular dataset do not fully coincide. Also

while RDF permits only binary relations, natural language expressions can more freely relate one, two or more arguments.

In order to capture linguistically rich information about verbalizations of simple and complex elements of an ontology or dataset, for example specifying arguments as optional or capturing restrictions on the usage of verbalizations, lexical knowledge is needed. Moreover, this lexical knowledge should become part of the linked data cloud itself, in order to avoid that it has to be recreated by every application that wants to use it. The *lemon* model[1] [5] has been developed for exactly this purpose, i.e. to create a standard format for publishing lexica as RDF data that declaratively state how vocabulary elements defined in a given ontology or used in a given dataset are verbalized in a particular language. However, the creation of such lexica is costly and while it can be automated to some extent, as shown for example in [9], it is highly desirable to share these lexica in accordance with linked data principles, so that everybody can benefit from them.

As proof of concept for a lexical layer enriching the linked data cloud, we manually developed an English lexicon for the DBpedia ontology in *lemon* format. In this paper we describe the release of the first version of this lexicon, illustrate the methodology that has lead to its creation and summarize the challenges in creating such a resource. Finally, we give an outlook on future plans and invite NLP and Semantic Web researchers to use the lexicon, improve and extend it, and to create similar lexical resources for other domains.

## 2 Method and dataset description

The DBpedia ontology[2] currently comprises 359 classes and 1,775 properties. Since the manual creation of lexical entries is an effort-intensive process, our approach to creating a lexicon for the DBpedia ontology is an iterative one. The first step consists in covering all classes and those properties that are most frequent with respect to the number of occurrences in triples in the DBpedia dataset. Later, we will successively extend the lexicon to also cover the tail of less frequent properties, ideally with support from the community.

The first release of the English DBpedia 3.8 lexicon comprises lexicalizations of 354 classes and 300 properties. The covered classes are complete except for one abstract class (`PersonFunction`) and a few classes without any instances (e.g. `NoteworthyPartOfBuilding`). The covered properties comprise all those with more than 10,000 occurrences in the DBpedia dataset. Excluded were the Wikipedia-specific ones (e.g. `wikiPageRevisionID` and `thumbnail`), abstract properties (e.g. `leaderFunction`), and properties for which no straightforward lexicalization was found (e.g. the datatype property `strength`, relating a battle to the numerical sizes of the involved military units).

The first release of the lexicon thus covers 98 % of the classes and approximately 20 % of the properties. We plan to successively extend the current lexicon,

[1] `http://lemon-model.net`
[2] `http://dbpedia.org/Ontology`

so that a second release will cover all properties with at least 1,000 occurrences (752 properties, i.e. 42 %), and a third release will cover all properties with at least 100 occurrences (1,112 properties, i.e. 63 %).

The lexicon currently contains 1,217 entries (443 class lexicalizations and 774 property lexicalizations), which amounts to approximately 1.8 entries per ontology concept (1.3 per class and 2.4 per property). The distribution of the number of lexicalizations per entry is depicted in Figure 1, where the $x$-axis specifies the number of entries, up to 5, and the bars specify how many concepts have this number of lexicalizations.

All lexical entries were created manually by two of the authors, familiar with both DBpedia and *lemon*, partly by manually selecting the most frequent patterns from the WikiFramework [3,1] and BOA [2] pattern libraries. The main objective when devising entries was to provide a wide range of lexical variants, especially those that differ from the RDFS label and thus are most likely to be helpful for NLP applications. For example, for the object property `spouse` the lexicon lists as verbalizations the nouns *spouse of*, *wife of* and *husband of*, as well as the verb *marry* and its participle form *married to*; for the datatype property `elevation` it lists the nouns *elevation of*, *altitude of* and *height of*, the verbs *stand at* and *rise to*, as well as the adjective *high*.

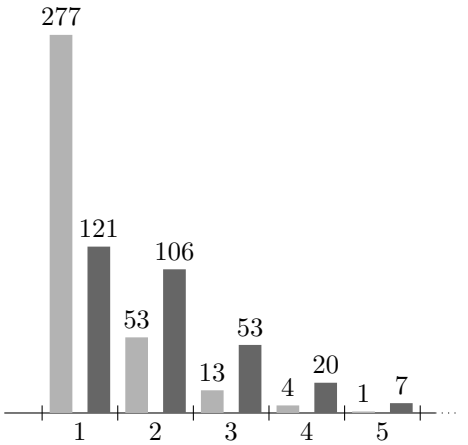

Fig. 1: Distribution of class lexicalizations (light grey bars) and property lexicalizations (dark grey bars)

Most of the entries were constructed using a domain-specific language[3] for common *lemon* design patterns [6]. Figure 2 gives some examples, specifying different frames (e.g. relational noun and state verb) together with a canonical form, a reference w.r.t. the ontology, and a mapping between semantic and syntact arguments.

Of all lexical entries, 54 could not be captured by the *lemon* design patterns but only by writing *lemon* RDF triples. This is mostly the case for constructions, such as *X has Y inhabitants* as verbalization of `population`, or *X consists to Y percent of water* as verbalization of `percentageOfAreaWater`, and for entries with a compound meaning, i.e. a sense that consists of several subsenses. An example is the verb *link to* as in *The Autostrada A19 links Palermo to Catania*, which refers to both `routeStart` and `routeEnd`. Both constructions and compound meanings in patterns will be subject of future developments.

[3] `https://github.com/jmccrae/lemon.patterns`

```
ClassNoun("mountain",dbpedia:Mountain)

RelationalNoun("author",dbpedia:writer,
      propSubj = PossessiveAdjunct,
      propObj  = CopulativeArg)

StateVerb("write",dbpedia:writer,
      propSubj = DirectObject,
      propObj  = Subject)

RelationalAdjective("based",dbpedia:headquarter,
      relationalArg = PrepositionalObject("in"))
```

Fig. 2: Example verbalizations of the class `Mountain` and the properties `writer` and `headquarter`.

In case a lexical entry verbalizes a concept that is not named in the ontology, it is defined in the lexicon. In total, the lexicon specifies 89 classes and 2 properties in addition to the DBpedia concepts. An example are the classes of all male or female persons, which are not part of the DBpedia ontology but can straightforwardly be defined as the restriction classes of all things with `gender` either `Male` or `Female`, e.g.:

```
lex:Female rdf:type       owl:Restriction ;
           owl:onProperty dbpedia:gender ;
           owl:hasValue   resource:Female .
```

These classes can be verbalized as *man* and *woman*. Additionally they can serve as domain or range restrictions, as e.g. in the verbalization *daughter of*, referring to the property `child` with its object restricted to the class `Female`, as shown in Figure 3.

```
ClassNoun("woman",lex:Female) with plural "women"

RelationalNoun("daugther",dbpedia:child,
   propSubj = PossessiveAdjunct,
   propObj  = CopulativeArg restrictedTo lex:Female)
```

Fig. 3: Example verbalizations of the additional lexical class `Female` and the property `child`.

Other common cases of compound meanings are properties for which not the verbalization of the property itself but rather the verbalization of the property together with a particular object is relevant. Examples are nationalities (e.g. `Dutch` being the class of all persons related to the resource `Netherlands` through the property `nationality`), occupations (e.g. `Surfer` being the class of

```
IntersectiveDataPropertyAdjective("extinct",
       dbpedia:conservationStatus,"EX")
IntersectiveDataPropertyAdjective("endangered",
       dbpedia:conservationStatus,"EN")
```

Fig. 4: Example verbalizations of conservation statuses

all persons related to the resource `Surfing` through the property `profession`), and religions (e.g. `Buddhist` being the class of all persons related to the resource `Buddhism` through the property `religion`). Another example is the datatype property `conservationStatus`, which characterizes a species as endangered, extinct, or the like. Since the conservation status code is not meaningful for non-experts, verbalizations such as *endangered* and *extinct*, as given in Figure 4, should be included in the lexicon.

Although our DBpedia lexicon currently covers only the top 30 % of all URIs in the DBpedia ontology, it can already prove useful for NLP applications. As a rough estimation, our lexicon provides verbalizations for 91 of the 104 different URIs used in the QALD-3[4] query set for question answering over linked data (i.e. for 95 %).

Note that the provided lexicon includes only schema but no instance data. For lexicalizations of the nearly 3 million individuals, for example the DBpedia lexicalization dataset [7] can be exploited, providing label alternatives for all resources, based on redirects, disambiguation links and Wikipedia anchor texts.

## 3 Release status and future plans

The first version of the DBpedia lexicon is released at `http://lemon-model.net/lexica/dbpedia_en/`, under the Creative Commons license CC BY 3.0[5]. It is accessible as open source on GitHub, at `https://github.com/cunger/lemon.dbpedia`, allowing others to improve and extend the lexicon as well as to port it to other languages. In the future, we plan to release the lexicon also under *lemon* source[6], a collaborative web interface for creating and editing lexica. Possibly even the DBpedia community could be involved in lexically enriching the ontology in multiple languages when editing the ontology wiki.

In order to get an idea of the coverage of the lexicon and its usefulness for NLP applications like question answering and natural language generation, we provide a demo at `purl.org/3dlt/demo`.

In future releases we will focus much more on a semi-automatic approach to creating lexical entries, as described in [9], in order to reduce the manual effort in the lexicon creation process. The manually created lexicon can then serve as gold standard for evaluating this and also other approaches to ontology lexicalization.

[4] `http://www.sc.cit-ec.uni-bielefeld.de/qald/`
[5] `https://creativecommons.org/licenses/by/3.0/`
[6] `http://monnetproject.deri.ie/lemonsource/`

Furthermore, we are creating a first version of a German and Spanish DBpedia lexicon, based on an automatic translation of the English lexical entries (cf. [4]) and a subsequent manual validation and correction step.

## 4 Conclusion

We described the first release of a manually constructed English lexicon for the DBpedia ontology in *lemon* format. The focus is on high quality entries, covering especially those lexicalizations that differ from the provided RDFS label, and those that verbalize complex constructs and thus cannot yet be handled by automatic lexicalization methods.

We hope that the first release of the DBpedia lexicon serves to prove the usefulness of rich lexical knowledge for NLP applications, and inspires NLP and Semantic Web researchers to improve and extend the lexicon, port it to other languages, as well as build and share lexica for other domains, thereby enriching the linked data cloud with a lexical layer.

**Acknowledgment** This work was partially funded within the EU projects PortDial (FP7-296170) and Monnet (FP7-248458).

## References

1. E. Cabrio, J. Cojan, A. Palmero Aprosio, B. Magnini, A. Lavelli, and F. Gandon. QAKiS: an open domain QA system based on relational patterns. In *Proc. of the 11th International Semantic Web Conference (ISWC 2012), demo paper*, 2012.
2. D. Gerber and A.-C. Ngonga Ngomo. Bootstrapping the linked data web. In *Proc. of the 10th International Semantic Web Conference (ISWC)*, 2011.
3. R. Mahendra, L. Wanzare, B. Magnini, R. Bernardi, and A. Lavelli. Acquiring relational patterns from Wikipedia: A case study. In *Proc. of the 5th Language and Technology Conference*, 2011.
4. J. McCrae, M. Espinoza, E. Montiel-Ponsoda, G. Aguado de Cea, and P. Cimiano. Combining statistical and semantic approaches to the translation of ontologies and taxonomies. In *Proc. of the Fifth workshop on Syntax, Structure and Semantics in Statistical Translation (SSST-5)*, 2011.
5. J. McCrae, D. Spohr, and P. Cimiano. Linking lexical resources and ontologies on the semantic web with lemon. In *Proc. of the 8th Extended Semantic Web Conference (ESWC)*. Springer, 2011.
6. J. McCrae and C. Unger. Design patterns for engineering the ontology-lexicon interface. In Paul Buitelaar and Philipp Cimiano, editors, *Multilingual Semantic Web*. Springer, to appear.
7. P.N. Mendes, M. Jakob, and C. Bizer. Dbpedia for nlp: A multilingual cross-domain knowledge base. In *Proc. of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.
8. A. Reymonet, J. Thomas, and N. Aussenac-Gilles. Modelling ontological and terminological resources in OWL DL. In *Proc. of OntoLex07*, volume 7, 2007.
9. S. Walter, C. Unger, and P. Cimiano. A corpus-based approach for the induction of ontology lexica. In *Proc. of the 18th International Conference on Application of Natural Language to Information Systems (NLDB 2013)*, 2013.