A Vocabulary Independent Generation Framework for DBpedia and beyond

Ben De Meester[1], Anastasia Dimou[1], Wouter Maroy[1], Dimitris Kontokostas[2], Ruben Verborgh[1], Jens Lehmann[3,4], Erik Mannens[1], and Sebastian Hellmann[2]

[1] Ghent University – imec – IDLab,
Department of Electronics and Information Systems, Ghent, Belgium
{firstname.lastname}@ugent.be
[2] Leipzig University – AKSW/KILT, Leipzig, Germany
{lastname}@informatik.uni-leipzig.de
[3] University of Bonn, Smart Data Analytics Group, Germany
jens.lehmann@cs.uni-bonn.de
[4] Fraunhofer IAIS, Sankt Augustin, Germany
jens.lehmann@iais.fraunhofer.de

Abstract. The DBpedia Extraction Framework, the generation framework behind one of the Linked Open Data cloud's central hubs, has limitations which lead to quality issues with the DBpedia dataset. Therefore, we provide a new take on its Extraction Framework that allows for a sustainable and general-purpose Linked Data generation framework by adapting a semantic-driven approach. The proposed approach decouples, in a declarative manner, the extraction, transformation, and mapping rules execution. This way, among others, interchanging different schema annotations is supported, instead of being coupled to a certain ontology as it is now, because the DBpedia Extraction Framework allows only generating a certain dataset with a single semantic representation. In this paper, we shed more light to the added value that this aspect brings. We provide an extracted DBpedia dataset using a different vocabulary, and give users the opportunity to generate a new DBpedia dataset using a custom combination of vocabularies.

Keywords: DBpedia, FnO, Generation, Linked Data, RML

1 Introduction

The *DBpedia Extraction Framework* (DBpedia EF) extracts raw data from Wikipedia and makes it available as Linked Data, forming the well-known and broadly used DBpedia dataset [6]. The majority of the DBpedia dataset is derived through *Wikipedia infobox templates*, after being annotated by the *DBpedia ontology*[5] [6]. The *rules* describing the DBpedia dataset generation from Wikipedia are executed by the DBpedia EF, defined by a worldwide crowd-sourcing effort, and maintained via the *DBpedia mappings wiki*[6]. Even though DBpedia is one of the central

[5] http://dbpedia.org/ontology/
[6] http://mappings.dbpedia.org/index.php/Main_Page

interlinking hubs in the Linked Open Data cloud [9], its generation framework has limitations reflected on the generated dataset [8, 10].

A major issue is that other schema(s) than the DBpedia ontology cannot be used to annotate Wikipedia pages. The DBpedia EF functions only with the DBpedia ontology, e.g., the predicate depends on the ontology term used for a certain attribute of an infobox. This occurs because the DBpedia EF selects the corresponding parser based on where the mapping template is used and which ontology term is selected, e.g., the *dbo:date* triggers the *Date parser*.

Thus, if an ontology term is not added to the DBpedia ontology, it cannot be used. For instance, no other predicate than the *dbo:location* may be used to indicate an entity's location because no triples will be generated.

Other vocabularies, such as the *schema.org*[7] vocabulary, cannot be used unless they are imported into the DBpedia ontology, or the DBpedia EF is adjusted[8], because, otherwise, it will not recognize the vocabulary's properties.

Similarly, depending on the mapping template and ontology term (predicate) which are used, a different data type can be assigned. For instance, depending on which predicate is used, the area in square kilometers generates an *xsd:double* but also a DBpedia datatype (*dbo:areaTotal*) that depends on the used predicate.

In this work, we use the semantic general-purpose and more sustainable framework that replaces the current DBpedia EF, which decouples extraction, transformations and mapping execution from the DBpedia EF, and enables generating high quality Linked Data that is not limited to the DBpedia use case, as presented in detail by Maroy et al. [7]. We specifically demo how this work enables us to easily make both small and large schema-level changes to the generated DBpedia data without influencing the remainder of the generation process. The demo is available at `https://rmlio.github.io/dbpedia-ef-schema-demo/`.

2 Sustainable Linked Data Generation

The current DBpedia EF is coupled and custom, which hampers maintenance and limits flexibility with respect to mapping, transformation, and used schema [7]. The mapping rules are a custom solution coupled to the DBpedia EF and ontology. Similarly, the data transformations are hard-coded, executed at different places within the DBpedia EF, and coupled with the DBpedia ontology.

To alleviate its limitations, the following requirements are proposed [7]: (i) **Declarative mapping rules** covering all generated RDF triples, and the underlying implementation can interpret them in each case, whether they refer to *schema* or *data* transformations [2]. (ii) **Decoupled extraction, transformation, and mapping** allowing different extraction strategies, transformation libraries, or mapping rules without requiring adjustments to the underlying implementation [2]. (iii) **A vocabulary independent solution** to annotate the extracted data values, independently of the preferred vocabulary. (iv) **Machine-**

[7] `http://schema.org`
[8] As done for `dcterms` (`http://purl.org/dc/terms/`) and `foaf` (`http://xmlns.com/foaf/0.1/`)

processable mapping rules allow assessment not only for syntax, but also for schema validation [3] or automated mapping rules generation [5].

To address these requirements, we developed a solution that fulfills the aforementioned requirements built on the RDF Mapping Language (RML) [4]. RML performs the schema transformations, and is aligned with FnO [1], that performs data transformations. Both schema and data transformations are thus covered using declarative machine-processable rules, instead of coupled, while the wikitext extractor is a separate module, allowing for a decoupled architecture. Most importantly though, the rules are independent of the vocabulary used.

3 Demo: Interchanging Schemas

The mapping rules which are described in RML, are RDF triples themselves. Thus, they can be updated – automatically or not – and other semantic annotations can be applied or other datasets can be generated from Wikipedia. Taking advantage of this and relying on the DBpedia mapping rules and the alignment of DBpedia ontology with schema.org[9], we translated the RML mapping rules for DBpedia to use schema.org and generate another RDF subgraph. More specifically, within the original mapping document, predicates and classes from the DBpedia ontology were replaced by predicates and classes from schema.org. No further changes were required, neither for the mapping rules, nor for the data transformations.

We executed a new extraction which was done over all 16,244,162 pages in the English DBpedia that contained articles, templates, media/file descriptions, and primary meta-pages. 191,288 Infobox_persons were found and 1,026,143 RDF triples were generated. Indicatively, 179,037 RDF triples were generated with `schema:name` property, 54,664 with `schema:jobTitle`, 23,751 with `schema:nationality`, 144,907 with `schema:birthPlace`, and 139,488 with `schema:birthDate`. The RDF dataset is available at `http://mappings.dbpedia.org/person_schema.dataset.ttl.bz2`, and can be interactively queried and compared with the original DBpedia dataset on `https://rmlio.github.io/dbpedia-ef-schema-demo/`.

Furthermore, at `https://rmlio.github.io/dbpedia-ef-schema-demo/`, we provide an interactive Web application that allows users to easily apply small and large changes in the DBpedia mapping document to change the schema of the resulting data. Presets are available to easily switch between annotations using the DBpedia ontology or schema.org, but users are encouraged to make their own changes and create hybrid solutions, or even use completely different ontologies and vocabularies. Users can trigger the generation of RDF data based on their applied changes, to review their adjustments. The application demonstrates that changes in the schema remain localized, e.g., changing the predicate does not influence which data type is used or which function is executed, i.e., schema transformations are decoupled from vocabulary and data transformations.

In this demo paper, we showcase a generic and semantics-driven approach to improve the Linked Data generation of the current DBpedia EF, as described in detail in [7], and provide a proof-of-concept to show the extended possibilities with

[9] `http://schema.org`

respect to schema transformations and how their changes remain decoupled from the remainder of the generation framework. Most importantly, the generation occurs independently of the vocabulary used to semantically annotate the RDF dataset. This is evident by providing a new DBpedia dataset that has all person resources mapped into RDF, however, using schema.org instead of the DBpedia ontology. This dataset can be interactively compared to the original DBpedia dataset. Furthermore, an interactive application allows users to make small and large changes in the schema when generating DBpedia data. They can easily switch between different vocabularies used to generate DBpedia data, and create custom schema transformations.

References

1. B. De Meester, A. Dimou, R. Verborgh, E. Mannens, and R. Van de Walle. An Ontology to Semantically Declare and Describe Functions. In *The Semantic Web: ESWC 2016 Satellite Events*, volume 9989 of *LNCS*, pages 46–49. Springer, 2016.
2. B. De Meester, W. Maroy, A. Dimou, R. Verborgh, and E. Mannens. Declarative data transformations for Linked Data generation: the case of DBpedia. In *The Semantic Web – Latest Advances and New Domains (ESWC 2017)*. Springer, 2017.
3. A. Dimou, D. Kontokostas, M. Freudenberg, R. Verborgh, J. Lehmann, E. Mannens, S. Hellmann, and R. Van de Walle. Assessing and Refining Mappings to RDF to Improve Dataset Quality. In *The Semantic Web – ISWC 2015*, volume 9367 of *LNCS*, pages 133–149. Springer, 2015.
4. A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. RML: A Generic Language for Integrated rdf Mappings of Heterogeneous Data. In *Proceedings of the 7th Workshop on Linked Data on the Web*, volume 1184 of *CEUR Workshop Proceedings*, 2014.
5. P. Heyvaert, A. Dimou, A.-L. Herregodts, R. Verborgh, D. Schuurman, E. Mannens, and R. Van de Walle. RMLEditor: A Graph-based Mapping Editor for Linked Data Mappings. In *The Semantic Web – Latest Advances and New Domains (ESWC 2016)*, volume 9678 of *LNCS*, pages 709–723. Springer, 2016.
6. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Sem Web*, 2015.
7. W. Maroy, A. Dimou, D. Kontokostas, B. De Meester, R. Verborgh, J. Lehmann, E. Mannens, and S. Hellmann. Sustainable linked data generation: The case of DBpedia. In *Proceedings of the 16th International Semantic Web Conference: In-Use Track*, Vienna, Austria, Oct. 2017. Springer.
8. H. Paulheim and A. Gangemi. *Serving DBpedia with DOLCE – More than Just Adding a Cherry on Top*, pages 180–196. Springer International Publishing, 2015.
9. M. Schmachtenberg, C. Bizer, and H. Paulheim. *Adoption of the Linked Data Best Practices in Different Topical Domains*, pages 245–260. Springer, 2014.
10. A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann. User-driven Quality Evaluation of DBpedia. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 97–104. ACM, 2013.