Fourth Summer Datathon on Linguistic Linked Open Data

Tijana Radović, Ranka Stanković

Professional paper

Fourth Summer Datathon on Linguistic Linked Open Data

UDC 81'322.2:004.822
DOI 10.18485/infotheca.2023.23.1.6

Tijana Radović
tijana.n.radovic@gmail.com
University of Belgrade
Belgrade, Serbia

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

ABSTRACT: The 4th Summer Datathon on Linguistic Linked Open Data (SD-LLOD-22) was held in Spain, in Cersedilla near Madrid, in May 2022, and organized by the COST Action NexusLinguarum. The school gathered interested researchers, academics, students who wanted to acquire and/or expand their knowledge in the field of linguistic linked data science. During the school, a spectrum of topics from the field of linked data was presented, from various ontologies, through document integration, annotation and natural language text processing tools based on linked data and RDF using Web Annotation and NIF (*NLP Interchange Format*) standards, to guidelines for generating RDF and publishing language resources as well as providing insight into the importance of using linked data in lexicography and terminology. The organizer of the school was the Polytechnic University of Madrid and the school had over 50 participants from both academia and industry.
KEYWORDS: linguistic linked open data, sentiment analysis, linked data, RDF.

PAPER SUBMITTED: 28 November 2022
PAPER ACCEPTED: 30 March 2023

1 Introduction

As part of the NexusLinguarum COST action, the 4th Summer Datathon on Linguistic Linked Open Data (SD-LLOD-22) was held in Spain, in Cercedilla, near Madrid, in the period from 05/30/2022 to 06/03/2022 at the Polytechnic University of Madrid (*Universidad Politécnica de Madrid*). COST action

NexusLinguarum[1] is funded by the European Union to achieve synergy between linguists, computer scientists, terminologists, language experts and other interested researchers across Europe, and to explore and expand the field of linguistic linguistic linked data science (Dojchinovski et al. 2021).

2 Program and organization of the school

The content of the *SD-LLOD-22* school consisted of lectures in the field of linked data intended for the participants who want to get a theoretical basis and familiarize themselves with the subject matter and the methodology used in this field. Afterwards, workshops were held in which concrete practical problems from different areas of data science were solved, with the main goal of providing participants from industry and academia with practical knowledge in the field of linked data applied to linguistics. And finally, through practical work on the realization of mini-projects, the participants worked in groups. Through their own mini-projects, they applied what they had learned including generating and/or using linguistically related data with the help of one of several available mentors who are experts in specific topics. The ultimate goal was to enable participants to transform existing linguistic data and publish it as linked data on the web and to develop applications that use linguistic linked data.[2] Similar events have been organized before: in 2015 and 2017 in Sersedilla (Spain) and in 2019 in Dagstul (Germany).

Various language resources such as dictionaries, corpora and knowledge bases are widely and openly available through linked data technology. However, there is still a large number of disconnected resources that are stored in heterogeneous formats and with different representation schemes. In addition, these resources do not have standard ways of programming data access API(*Application Programming Interface*). Representation of language resources as linked data has numerous advantages such as aggregation and integration of language resources using a common data model RDF (*Resource Description Framework*), explicit linking, publishing and searching in a standardized way (using SPARQL), improved discovery of datasets and services, as and using common vocabularies to represent language content and metadata. In short, linked data enables easier search of language resources, their availability, interpretability and reusability (Gracia et al. 2022a).

1. 4th Summer Datathon on Linguistic Linked Open Data (SD-LLOD-22)
2. Information about the event and links to resources can be found at Datathon2022

During the school, the participants: 1) generated and published their own linguistic linked data from already existing data sources, 2) applied the principles of linked data and semantic web technologies in the field of linguistic resources, 3) used some of the most important models used to represent linguistic linked data data, especially *OntoLex-Lemon* (McCrae et al. 2017), 4) became familiar with natural language text processing flows and applications based on linked data, 5) gain insight into the potential benefits of linked data and the possibilities of their application for specific use cases.

The following topics were presented during the school: ontologies and related data with an emphasis on the *OntoLex-Lemon* model as a lexicon model for ontologies that supports RDF (*Resource Description Framework*) and OWL (*Web Ontology Language*), two of the three fundamental technologies used within the semantic web (Gracia et al. 2022b). Document integration, annotation, and natural language text processing tools based on linked data and RDF using Web Annotation and NIF (*NLP Interchange Format*) standards are then presented. Then the guidelines for generating RDF and publishing language resources as well as the importance of using linked data in lexicography and terminology are presented. At the end, a review is given of the overall importance of the use and application of linguistically related data to the potential benefits of linked data and the possibilities of their application for specific use cases.

3 *Sentimientos* – one of the mini-projects realized during SD-LLOD-22

As it was mentioned before, one aspect of the school involved practical work in groups through which participants applied what they had learned through their own mini-projects, including the generation and/or use of linguistically related data. This independent practical work is designed so that participants propose their own "mini-projects" based on their interests, which include the conversion of existing datasets into linked data, with mini-projects for under-resourced languages particularly encouraged. The representatives of Serbia at this event were PhD students of intelligent systems at the University of Belgrade: Tijana Radović and Miloš Košprdić. Together with colleague Martin Alejandro Chaya from the University of Madrid and mentor Sin Ahmadi from the University of Bologna, they worked on the *SD-LLOD-22 Sentimientos* project which aimed to enrich, convert and publish lexicographic data with word polarity information in a linked data format. The *Senti-Pol-sr* lexicon (Stanković et al. 2022) was used for the Serbian language, and the *PolArt*

lexicon (Klenner, Fahrni, and Petrakis 2009) for the German language. The structure of both lexicons is quite simple: a lemma and associated polarity expressed through positive and negative columns with discrete values 0 and 1.

The project consisted of three parts:

1) Starting from the list of lemmas, lexicographic data was collected from the *Dbnary* database (Sérasset 2012) using SPARQL queries. An ontology template was created: a schema of lexical entries consisting of a lexical entry (*LexicalEntry*), a type of speech (*partOfSpeech*), a lemma (*'lemma_label'*), a lexical sense (*lexicalSense*) and a polarity (*hasPolarity*), which can be seen in the following code (the parts that change are marked in blue). Using the presented template, each record from the input dictionaries was further transformed into RDF form.

```
rdf_template = """
<:le_uri> a ontolex:lexicalEntry;
   lexinfo:partOfSpeech <:pos_uri>;
   rdfs:label 'lemma_label'@language_tag;
   ontolex:lexicalSense :senses_uris.

<:lemma_opinion> marl:describesObject <:le_uri>;
   marl:hasPolarity polarity.
"""
```

2) To produce the so-called *ttl* file, the data is augmented using SPARQL queries against the *Dbnary* lexicon to retrieve the lexical entry, word type and meaning for specific words from the input dictionaries. The SPARQL query for supplementing existing data is shown below:

```
query = (
  "\n"
  " PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>\n"
  " PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>\n"
  " PREFIX lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>\n"
  " \n"
  " SELECT ?le ?pos ?sense \n"
  " WHERE {\n"
  " ?le a ontolex:LexicalEntry .\n"
  " ?le rdfs:label \"xxxxx\"@<language2>.\n"
  " ?le lexinfo:partOfSpeech ?pos.\n"
  " ?le ontolex:sense ?sense\n"
  " }"
```

```
"")
```

3) The query is forwarded to the *Dbnary* access point, in this case as follows:

```
sparql = SPARQLWrapper( "http://kaiko.getalp.org/sparql")
sparql.setReturnFormat(JSON)

def run_query(word, language, language_):
    word_query = query.replace("xxxxx", word)
    word_query = word_query.replace("<language2>", language)
    sparql.setQuery(word_query)
    return sparql.queryAndConvert()["results"]["bindings"]
```

An example of a lexical entry containing lemma, part of speech and polarity of sentiment would look like this in the transformation presented in Figure 1.

```
### word: hrana
<http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1>
  a ontolex:lexicalEntry;
  lexinfo:partOfSpeech <http://www.lexinfo.net/ontology/2.0/lexinfo#noun>;
  rdfs:label 'hrana'@sr;
  ontolex:lexicalSense
  <http://kaiko.getalp.org/dbnary/ell/__ws_1__srp__hrana__ουσιαστικό__1>.

<http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1/opinion>
  marl:describesObject
  <http://kaiko.getalp.org/dbnary/ell/_srp__hrana__ουσιαστικό__1>;
  marl:hasPolarity marl:positive.
```

Figure 1. An example of a lexical entry.

In this mini-project, the aforementioned Serbian and German lexicons of sentiment and *Dbnary* were used, but the approach can be extended to other types of lexicons. In order to integrate the produced tools into the user application, the produced data were stored in the *GraphDB* database[3] and

3. GraphDB

a small application was created over them using a Java script program for searching the stored data.

The input and output, as well as the developed code, are publicly available,[4] while an additional version enhanced with part of speech from the *Serbian Morphological Dictionary* (Krstev 2008) and the *Leximirka* lexical database (Stanković et al. 2018). The *Senti-Pol-sr.ttl* file for direct access and download is available at the site of the Association of Language Resources and Technologies – JeRTeh,[5] while SPARQL search is available at the JeRTeh's SPARQL end point implemented using the *Fuseki* tool.[6]

4 Conclusion

The Fourth Summer Datathon on Linguistic Linked Open Data was valuable for us as participants, as an opportunity to connect and share knowledge, ideas and experiences with researchers with common interests from both academia and industry. The content was diverse, with topics adapted to the knowledge of the participants in different sections, so that we were able to expand previously acquired knowledge and obtain guidelines for future research. Lectures of the theoretical type helped us to better understand the methodology and concepts of linking, as well as the schemas and standards of data representation. At the workshops, by working on specific problems, we got acquainted with different technologies in the field of linked data. The knowledge and skills gained during the 5-day duration allowed us to generate and publish our own linguistic related data from already existing data sources, which will enable us to conduct further research in this direction: generation of similar data sets, enrichment and expansion of the published lexicon. Also, we got acquainted with natural language text processing flows and applications based on linked data and became aware of the possibilities of their application. The work on the *Sentimientos* mini-project contributed to the enrichment of resources in the Serbian language by converting and publishing a word list with associated information about part of speech and polarity in the format of linked data (in the form of a *ttl* file and on an access point), and can serve as a resource for future research related to determining the sentiment of the text, but also for publishing linguistic data in a similar way. Linking the lexicon with usage examples that would be published in NIF format is one of the first tasks that we will tackle in the coming period.

4. SD-LLOD-22 Sentimientos
5. http://llod.jerteh.rs
6. http://fuseki.jerteh.rs//dataset/Senti-Pol-sr/query

Acknowledgment

The presented work is supported by the the *COST Action CA18209 – NexusLinguarum "European network for Web-centred linguistic data science"*. The authors would like to thank Miloš Košprdić for his participation in dataset preparation.

References

Dojchinovski, Milan, Julia Bosque Gil, Jorge Gracia, and Ranka Stanković. 2021. "EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School." *Infotheca* 21 (1). https://doi.org/10.18485/infotheca.2021.21.1.7.

Gracia, Jorge, Patricia Martín Chozas, Anas Fahad Khan, Christian Chiarcos, Elena Montiel Ponsoda, Sina Ahmadi, Thierry Declerck, et al. 2022a. *SD-LLOD-22 Course Slides,* October. https://doi.org/10.5281/zenodo.7197674.

Gracia, Jorge, Patricia Martín Chozas, Anas Fahad Khan, Christian Chiarcos, Elena Montiel Ponsoda, Sina Ahmadi, Thierry Declerck, et al. 2022b. *SD-LLOD-22 Materials,* October. https://doi.org/10.5281/zenodo.7197696.

Klenner, Manfred, Angela Fahrni, and Stefanos Petrakis. 2009. "Polart: A robust tool for sentiment analysis." In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, 235–238.

Krstev, Cvetana. 2008. *Processing of Serbian. Automata, texts and electronic dictionaries.* Faculty of Philology, University of Belgrade.

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Sérasset, Gilles. 2012. "Dbnary: Wiktionary as a LMF based Multilingual RDF network." In *Proc. of the 8th Int. Conf. LREC'12*, edited by Nicoletta Calzolari et al. ELRA. http://www.lrec-conf.org/proceedings/lrec2012/index.html.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, 31–38.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic Dictionaries – from File System to *lemon* Based Lexical Database." In *Proc. of the 11th Int. Conf. LREC 2018 - W23 6th Workshop on Linked Data in Linguistics : Towards Linguistic Data Science (LDL-2018), Miyazaki, Japan, May 7-12, 2018*, edited by John P. et al McCrae. ELRA. http://lrec-conf.org/workshops/lrec2018/W23/pdf/3_W23.pdf.