Undefined 0 (0) 0
IOS Press

Querying the LLOD Cloud for Linguistic Resources

Richard Littauer [a,b], Steven Moran [c,d], and Boris Villazon-Terrazas [e]
[a] *Department of Intelligent Computer Systems, University of Malta, Msida, MSD2080, Malta*
[b] *Computational Linguistics Department, Saarland University, Saarbrücken, 66121, Germany*
E-mail: littauer@coli.uni-saarland.de
[c] *Department of Linguistics, University of Zürich, Plattenstrasse 54, CH-8032 Zürich, Switzerland*
[d] *Research Unit Quantitative Language Comparison, Ludwig Maximilian University, Geschwister Scholl Platz 1, D-80539 Munich, Germany*
E-mail: steve.moran@lmu.de
[e] *Intelligent Software Components, iSOCO, S.A., Av. del Partenon 16-18, Madrid, Spain*
E-mail: bvillazon@isoco.com

Abstract. The Semantic Web offers a unique opportunity to query large amounts of disparate data that otherwise exist in fragmented databases. The Open Linguistics Working Group has been developing a cloud of RDF resources and ontologies that conform to the Linked Open Data paradigm, enabling language researchers and computational linguists to utilize rich linguistic data in the Semantic Web for analysis. Here, we present a SPARQL endpoint to the Linguistics Linked Open Data (LLOD) cloud. We illustrate how to query for language resources within the cloud, with example query results for language resources from multiple databases, which previously would have had to have been collected individually and arduously. Our aim is expedite the use and adoption of the LLOD cloud by language researchers and Semantic Web enthusiasts.

Keywords: Semantic Web, Linked Data, LLOD, Linguistics, Typology, Language Resources

1. Introduction

The amount of linguistic resources available on the Web has grown considerably in recent years. This is true of both data resources, such as lexical data, multimedia recordings and annotated corpora, as well as for computational tools cleaning and analysis, such as chunkers, part of speech taggers and parsers. However, limited interoperability between data formats and data licensing constraints are major obstacles for data use, reuse and sharing. Representing data so that they conform to the Semantic Web framework fosters interoperability of resources, as well as providing an infrastructure with a vibrant development community that can be used to query these resources. Linked Data refers to Semantic Web framework best practices for publishing and connecting structured data. One initiative to share openly available data published in Linked Data is called Linked Open Data.

The Linguistic Linked Open Data (LLOD) cloud[1] is a sub-cloud of the Linked Open Data cloud,[2] consisting of data from databases of linguistic corpora and metadata that conforms to the Linked Open Data paradigm [1]. This cloud is being developed by the Open Linguistics Working Group (OWLG) [3], an open community that aims to promote open data in linguistics and that facilitates communication between researchers from different scientific disciplines and communities. The LLOD contains data from dozens of publicly available online databases that have been converted to RDF. Here, we briefly outline the technolo-

[1] http://linguistics.okfn.org/resources/llod/
[2] http://richard.cyganiak.de/2007/10/lod/

gies behind the LLOD (described at length elsewhere, cf. [2,4]), and, for the first time we discuss accessing data in the LLOD through a SPARQL endpoint set up during the Multilingual Linked Open Data for Enterprises workshop. We provide a couple of examples of how to use the endpoint to query across resources linked in the LLOD. The first query displays all resources within the cloud that have information related to a language-specific ISO 639-3 three letter language name identifier. Once data sources that contain information about a given language are identified, each data source can then be further queried for detailed information regarding that given language. Our second example query identifies the typological features of a given language that are available in the World Atlas of Language Structures (WALS) [6]. We conclude by briefly discussing the possible applications of querying the LLOD cloud for linguistic analysis and its potential use to language researchers.

2. The Semantic Web

The representation formalisms and technologies that make up the Semantic Web can be used to enable interoperability for language databases on a wide scale. The purpose of Linked Data is to enable structured data to be shared on the Web. The Linked Open Data paradigm sets out four rules for representation of web resources [1]:

1. referred entities should be designated by using Uniform Resource Identifiers (URIs),
2. these URIS should be resolvable over HTTP,
3. data should be represented by means of specific W3C standards (such as RDF),
4. and a resource should include links to other resources.

These rules facilitate data interoperability in that they require that entities are addressed in a globally unambiguous way (1), that they can be accessed (2) and interpreted (3), and that entities that are associated on a conceptual level are also physically associated with each other (4) [4]. An essential part of Linked Data is the Resource Description Framework (RDF), a language that was developed to represent information, such as metadata, about resources on the Web. In an RDF graph, information is expressed with subject-predicate-object triples. Each component of the triple is encoded by a URI, making it globally unambiguous on the Web. These resources can also link to other resources in a standardized way, so that various RDF graphs can share links and be combined into larger graphs. SPARQL is the query language for RDF data and it also consists of triple patterns, with additional optional patterns and conjunctions, disjunctions, etc. [7]. One tool that instantiates the SPARQL protocol is the SPARQL endpoint. A SPARQL endpoint is a service that supports querying RDF data from a single graph over HTTP and it may also be used to query multiple distributed RDF graphs via endpoints. Thus attain federated query across multiple resources.

3. The LLOD cloud

The Linguistic Linked Open Data (LLOD) cloud represents a very small but focused portion of the Semantic Web. It is specifically a so-called sub-cloud of the Linked Open Data cloud. The LLOD cloud is maintained by the Open Linguistics Working Group (OWLG),[3] which has three main goals for promoting openness in Linguistics:

1. Promoting the idea of open linguistic data and resources
2. Developing the means for the representation of open data
3. Encouraging the exchange of ideas across different disciplines

Building an interoperable, linked data cloud is directly in line with these aims. The umbrella organization for the OWLG and other working groups, the Open Knowledge Foundation defines *openness* as: "A piece of content or data [that] is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and share-alike."[4] With this principle in mind, data from different resources and typological databases were converted into RDF and added to the LLOD cloud. Figure 1 provides an illustration of the current LLOD cloud and its contents.[5]

The OWLG stipulates the following criteria for adding new resources to the LLOD cloud:

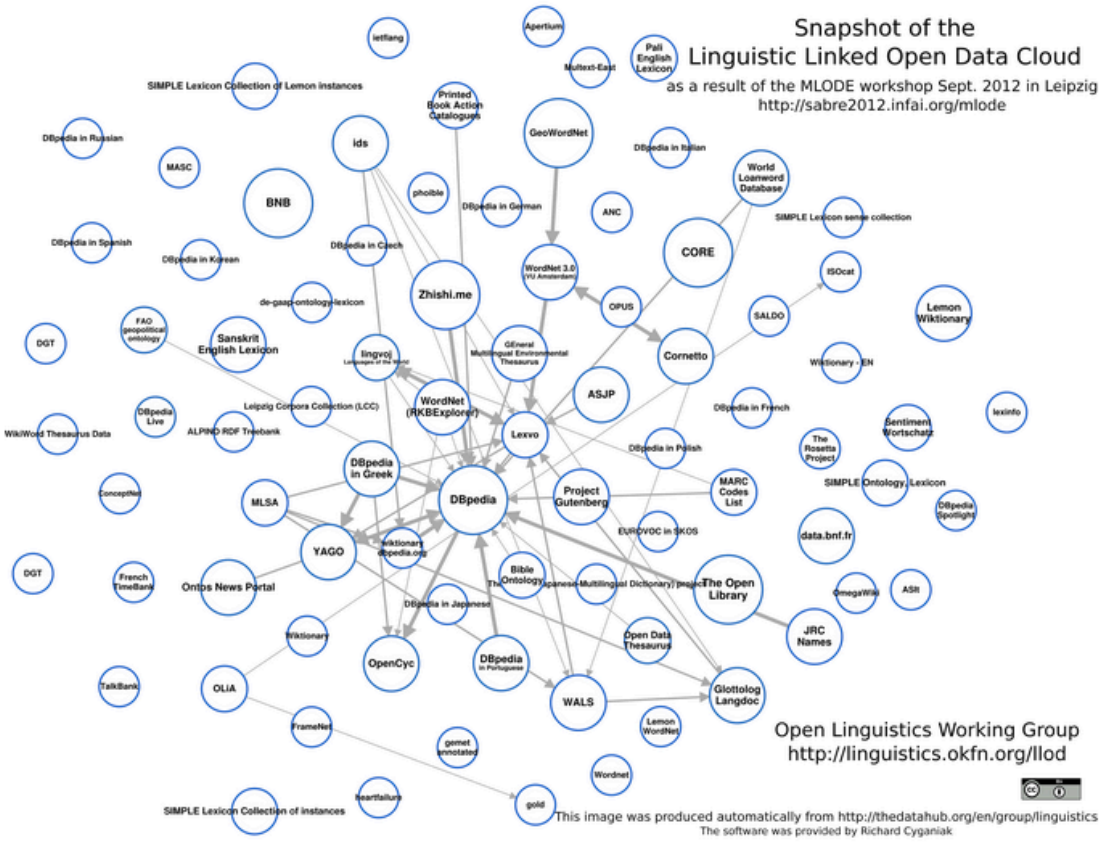
1. data is resolvable through HTTP
2. it is provided as RDF
3. it contains links to another data set in the diagram

[3] `http://linguistics.okfn.org/`

[4] `http://opendefinition.org`

[5] This image was generated with software written by Richard Cyganiak. For an illustration of the entire LOD cloud, see: `http://lod-cloud.net`.

Fig. 1. The LLOD Cloud



4. the entire data set must be available

At the time of writing, the cloud is in *draft* status, meaning that several of the resources may point only to resource metadata, even though the authors of these resources are committed to creating new links between to other linguistic datasets in the cloud. There are already many large and broad datasets that are linked in and across the LLOD cloud, including DBpedia, RDF versions of WordNet, Cornetto (Dutch WordNet), OpenCyc, and the Open Data Thesaurus. Linguistic-specific resources include data from typological databases like the World Atlas of Language Structures [6]. There are metadata repositories, such as Glottlog/Langdoc, Lexvo and lingvoj that provide pertinent information about languages, such as bibliographic references for source materials, ISO codes for language name identifiers and language families, alternative language name indices, and information such as where the languages are spoken. Lastly, there are ontological resources that describe grammatical features of languages and their relationships, such as the General Ontology of Linguistic Description [5], and for terminology resolution ISOcat.[6]

4. Querying the LLOD for language resources

A SPARQL endpoint for the LLOD was set up as part of the MLODE conference. This endpoint is located at: `http://mlode-sparql.nlp2rdf.org/sparql`.

In order to test the use of this endpoint, and in order to showcase how the LLOD can be queried and used in a non-theoretical capacity, we present the two queries below. These are by no means the only queries possible on the endpoint, nor do they represent the full nature of the LLOD – rather, these were judged to be of the most interest to linguistic researchers using online databases already.

[6] `http://www.isocat.org/`

Table 1
Query for resources with a given ISO 639-3 code

```
prefix wals: <http://wals.info/language/>
select distinct ?relation where {
wals:chr <http://purl.org/dc/terms/relation> ?relation .
}
```

Table 2
Result for query for resources with a given ISO 639-3 code

relation
http://www.llmap.org/maps/by-code/crw.html
http://www.ethnologue.com/show_language.asp?code=crw
http://en.wikipedia.org/wiki/ISO_639:crw
http://www.lexvo.org/data/iso639-3/crw
http://www.sil.org/iso639-3/documentation.asp?id=crw
http://multitree.org/codes/crw
http://scriptsource.org/lang/crw
http://www.language-archives.org/language/crw
http://odin.linguistlist.org/igt_urls.php?lang=crw
http://linguistlist.org/forms/langs/LLDescription.cfm?code=crw
http://www.glottolog.org/resource/languoid/id/chra1242

Using the SPARQL endpoint, we set ourselves the goal of devising SPARQL queries to identify all resources in the LLOD cloud that have data with regard to a specific ISO 639-3 unique language name identifier. ISO 639-3 identifiers are maintained by the Summer Institute of Linguistics[7] and consist of three-letter codes that refer to one of the existing 7000+ extant languages in the Ethnologue database.[8] This query is given in Table 1. The three-letter language code used is [crw], the language code for Chrau, a Vietnamese language spoken in Southeast Asia.

To use the endpoint, go to the link above and paste the query into the 'Query Text' box, and press 'Run Query'.[9] The query will then fire. The hyperlink of the loaded query can be used as a way to refer to that result without needing to re-enter the query for each iteration. The output of the query, at the time of writing, is given in Table 2. There are several resources that contain a variety of information on the Chrau language, including details about its geographical location (where speakers of Chrau live), its population, language family information, the system used to write the language, etc. As is clear, there are several datasets already available in the LLOD clod that can be used to gather more, related information about specific information.

Our second example query retrieves all of the information in WALS for a specific ISO 639-3 code. Along side ISO 639-3 codes, WALS also defines its own language name identifiers because its compilers have a different idea of what the set of mutually unintelligible language varieties are. For this reason, researchers often need to create wrappers to mine data several resources, such as the Ethnologue (the originator of language code), SIL (the gatekeeper of the current ISO 639-3 codes), and typology-specific databases and like WALS. It is of course possible to formulate a query that runs using the WALS language code, then finds the ISO 639-3 code, and retrieves all the information from other databases related to that code. The use of this sort of query for mining information about languages, which have various alternative names and even different 'standard' codes identifying those name, cannot be understated due to the amount of time and effort it saves the researcher, especially from search multiple different resources. A query that gathers WALS data is given in Table 3. The results of the query are given in Table 4. Here we have limited the results to 5 entries.[10]

[7] http://www.sil.org/iso639-3/

[8] A full list can be found here: `http://www.sil.org/iso639-3/download.asp`

[9] Uncheck the 'Default Data Set Name (Graph IRI)' in the top entry box by deleting `http://mlode.nlp2rd.org`, so that the box is empty.

[10] This can be done by appending LIMIT 5 after the closing bracket at the end of the code snippet.

Table 3
Result for query for resources with a given ISO 639-3 code

relation
http://www.llmap.org/maps/by-code/crw.html
http://www.ethnologue.com/show_language.asp?code=crw
http://en.wikipedia.org/wiki/ISO_639:crw
http://www.lexvo.org/data/iso639-3/crw
http://www.sil.org/iso639-3/documentation.asp?id=crw
http://multitree.org/codes/crw
http://scriptsource.org/lang/crw
http://www.language-archives.org/language/crw
http://odin.linguistlist.org/igt_urls.php?lang=crw
http://linguistlist.org/forms/langs/LLDescription.cfm?code=crw
http://www.glottolog.org/resource/languoid/id/chra1242

Table 4
Query for all information for a given ISO 639-3 code on WALS

```
prefix wals: <http://wals.info/language/>
select distinct  ?label ?descr ?ref ?area ?lat ?long ?genus
where
{
?s <http://purl.org/dc/terms/subject> wals:chr .
?s <http://wals.info/vocabulary/hasValue> ?value .
?value <http://purl.org/dc/terms/description> ?descr .
wals:chr <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?lat .
wals:chr <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?long .

wals:chr ?feature ?datapoint .
wals:chr rdfs:label ?label .
?datapoint <http://purl.org/dc/terms/references> ?ref .

?feature <http://purl.org/dc/terms/isPartOf> ?chapter .
?chapter <http://wals.info/vocabulary/chapterArea> ?area .
wals:chr <http://wals.info/vocabulary/hasGenus> ?genus.
wals:chr <http://wals.info/vocabulary/altName> ?name .
}
```

5. Conclusions

The two simple queries presented in our short report on tools gives readers an impression of some of the possibilities that are now available for querying resources in the LLOD cloud. As more datasets are converted to RDF and linked to the cloud, much more informational will be able for retrieval and analysis for language research and for commercial applications.

References

[1] T. Berners-Lee. Design issues: Linked data. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[2] C. Chiarcos, S. Hellmann, and S. Nordhoff. Linking linguistic resources: Examples from the Open Linguistics Working Group, this vol. p. 201-216.

[3] C. Chiarcos, S. Hellmann, S. Nordhoff, et al. The Open Linguistics Working Group. In *Proc. LREC 2012*, Istanbul, Turkey, May 2012.

[4] C. Chiarcos, S. Moran, P. N. Mendes, S. Nordhoff, and R. Littauer. Building a linked open data cloud of linguistic resources: Motivations and developments. In I. Gurevych and J. Kim, editors, *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer, to appear.

[5] S. Farrar and D. T. Langendoen. A Linguistic Ontology for the Semantic Web. *GLOT International*, 7:97–100, 2003.

[6] M. Haspelmath, M. Dryer, D. Gil, and B. Comrie, editors. *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich, 2008.

[7] E. Prud'Hommeaux and A. Seaborne. SPARQL query language for RDF. *W3C working draft*, 4(January), 2008.

Table 5
Results (LIMIT 5) for WALS for a given ISO 639-3 code

label	descr	ref	area	lat	long	genus
Chrau	The language has no morphologically dedicated second-person imperatives at all	Thomas 1971	Verbal Categories	10.75	107.5	http://wals.info/genus/bahnaric
Chrau	The prohibitive uses the verbal construction of the second singular imperative and a sentential negative strategy not found in (indicative) declaratives	Thomas 1971	Verbal Categories	10.75	107.5	http://wals.info/genus/bahnaric
Chrau	Adpositions without person marking	Thomas 1971	Verbal Categories	10.75	107.5	http://wals.info/genus/bahnaric
Chrau	Differentiation: one word denotes 'hand' and another, different word denotes 'finger' (or, very rarely, 'fingers')	Thomas 1971	Verbal Categories	10.75	107.5	http://wals.info/genus/bahnaric
Chrau	Identity: a single word denotes both 'hand' and 'arm'	Thomas 1971	Verbal Categories	10.75	107.5	http://wals.info/genus/bahnaric