Undefined 1 (2009) 1–5
IOS Press

Using the LLOD Cloud for Querying Linguistic Resources

Richard Littauer [a,b], Steven Moran [c], and Boris Villazon-Terrazas [d]
[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] *Research Unit Quantitative Language Comparison, Ludwig Maximilian University, Geschwister Scholl Platz 1, D-80539 Munich, Germany*
E-mail: bambooforest@gmail.com
[d] *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 for large amount of data which would otherwise exist only in fragmented databases. Recently, the Open Linguistics Working Group has been developing a sub cloud of ontologies which conform to the Linked Open Data paradigm with the intent to enable computational linguists to utilise the Semantic Web. Here, we present a SPARQL endpoint for the Linguistics Linked Open Data cloud. We showcase 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. It is hoped that the work presented here will expedite use of the Linguistics Linked Open Data cloud by researchers.

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

1. Introduction

The amount of computational linguistic resources available has grown considerably in recent years. This is true both of primary resources - audio corpora, dictionaries, ontologies - and secondary resources - parsers, segmenters, WordNets. However, limited interoperability and licensing constraints between primary resources are major obstacles for data use and reuse. Interoperability can be either structural, apparent when similar formalisms are used to access data, or conceptual, where the resources themselves share a common vocabulary [6]. Representing data so that they conform to the paradigms and languages of the Semantic Web fosters interoperability for resources, as well as providing an infrastructure with a vibrant development community that can be used to query these resources.

The Linguistics Linked Open Data (LLOD) cloud is a (sub-)cloud of the Linked Open Data cloud, consisting of databases of linguistic corpora or 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 dedicated towards creating and developing standards and linguistics resources for public use. The LLOD already contains dozens of publicly available online databases. Here, we briefly outline the technologies behind the LLOD (described at length elsewhere, cf. [2,4]), and, for the first time, provide a SPARQL endpoint to the LLOD, along with example queries of potential use to linguistics researchers. We show two queries with their results, the first displaying all of the databases within the LLOD cloud which have information related to the language-specific ISO code for an example language (each database can then be queried either individually or together). The second query gathers all of the ty-

pological data available through the endpoint from the World Atlas of Language Structures (WALS) [5] typological database, also for the same example language. We conclude by briefly discussing possible applications of querying the LLOD cloud.

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 Linked (Open) Data paradigm [1] sets out four rules for representation of web resources:

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 information integration, and thus, interoperability, in that they require that entities can be 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]

Following these rules, the Resource Description Framework (RDF) was laid down as a language meant to provide metadata for various resources. In RDF, information is expressed in triples (property, subject, and object), and RDF collections of data are represented by URIs, which are globally unambiguous in the web. This way, resources can link to each other, and an valid, coherent ontology can be more easily created.

SPARQL [7] is a standardised query language for RDF data. Inspired by SQL, SPARQL also uses a triple notation similar to RDF. It not only supports querying data from a single database accessible over HTTP - known as SPARQL endpoints - but also can query multiple databases from a single endpoint. It is this feature which allows the use of massive ontologies to be made out of several databases, creating a cloud of interoperable resources. RDF and SPARQL together are thus the main constituents of the Semantic Web.

3. The LLOD cloud

The Linguistics Linked Open Data cloud represents a portion of the Semantic Web network, and is specifically a sub cloud of the Linked Open Data cloud. It is maintained by the Open Linguistics Working Group[1], which has three main goals for promoting openness in Linguistics:

1. Promoting the idea of open linguistic 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 Open Knowledge Foundation, the umbrella organisation for the OWLG and other working groups, 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."[2] With this in mind, different databases within the Linked Open Data cloud have been marked out for inclusion within the LLOD. A diagram for the LLOD, developed by Cyganiak and Jentzsch, can be found at `http://lod-cloud.net` - this diagram is displayed in Fig. 1.

The following criteria must be met for a new linguistic resource to be included in the LLOD cloud:

1. The data is resolvable through HTTP,
2. it is provided as RDF,
3. it contains links to another data set in the diagram, and
4. the entire data set must be available.

. At the time of writing, the LLOD has *draft* status, meaning that several of the resources may point only to resource metadata, although each has been promised to be uploaded and linked to the LLOD in the near future. There are already many resources within the cloud, however, such as DBpedia, different RDF versions of WordNet, Cornetto (Dutch WordNet), OpenCyc, and the Open Data Thesaurus, metadata repositories like Lexvo and lingvoj. GOLD and ISOcat are currently available, although their license conditions are yet to be clarified.

4. Querying the LLOD for language resources

[1] `http://linguistics.okfn.org/`
[2] `http://opendefinition.org`

Fig. 1. The LLOD Cloud

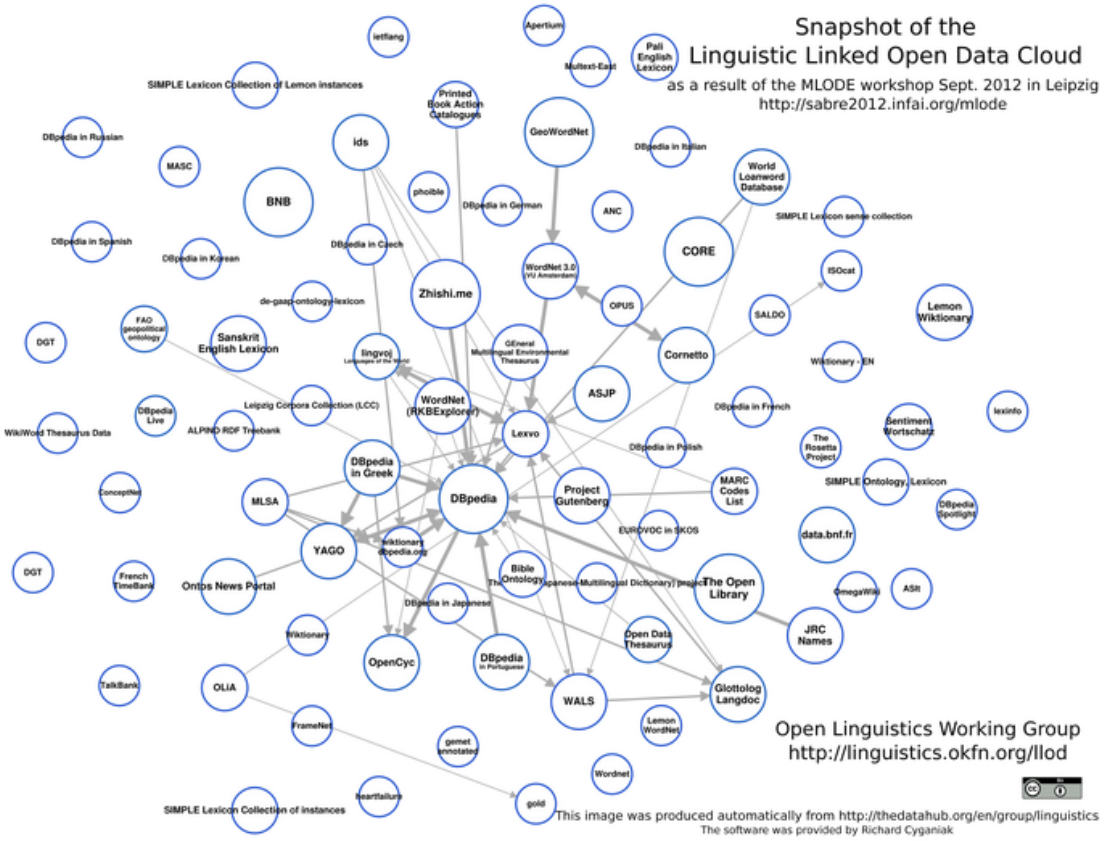

Table 1

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

A SPARQL endpoint for the LLOD was set up as part of the MLODE conference, and there are plans to see that it remains in public use. This endpoint can be found 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 wrote the two queries presented 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.

Using the SPARQL endpoint described above, we set ourselves the goal of devising SPARQL queries to identify all resources in the LLOD that have data with regard to a given ISO 639-3 unique language name identifier. ISO 639-3 identifiers are maintained by the Summer Institute of Linguistics[3], and consist of three-letter codes which refer to one of the existing 7000+ extant languages in the database.[4] This query can be seen in Table 1. The three-letter language code used is `chr`, the language code for Chrau, a Vietnamese language. This code was chosen at random.

To use the endpoint, go to the link above. Remove the 'Default Data Set Name (Graph IRI)' in the top entry box by deleting `http://mlode.nlp2rd.org`, so that the box is empty. Write your query into the 'Query Text' box, and press 'Run Query.' The query will then load. The hyperlink of the loaded query can be used as a way to refer to that result without needing to re-enter the query each iteration. The output of this query, as of the time of writing, can be seen in Table 2.

As is clear, there are several databases already available in the LLOD which can be used to gather more, related information. Notable among the available databases are: Multitree, which shows phylogenic relationships between languages; glottolog, which is a database of reference information for languages; ODIN, a database of language corpora automatically extracted from research papers around the web; Wikipedia; and LinguistList, which lists current researchers and reference works for each language.

The next query retrieves all of the information in WALS for a given ISO 639-3 code. WALS also uses their own language identifier, along with the ISO 639-3 code. For this reason, researchers often need to create wrappers to mine data bot from the Ethnologue, SIL, or Linguist databases and from WALS. It is hypothetically possible to create a query that would run using the WALS language code, find the ISO code, and then retrieve all information from other databases for that code. The use of this sort of query cannot be understated in the amount of time and effort saved. The query gathering WALS data can be seen in Table 3.

The results of the query in Table 3 can be seen in Table 4. Here, we have limited the results to 5 entries. (This can be done by appending LIMIT 5 after the closing bracket at the end of the code snippet.)

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

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

5. Conclusions

The queries presented here are a small subset of the possibilities of the LLOD cloud. As more databases are added, more informational will be able to bee retrieved that can be used in linguistic research or for commercial uses.

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

[6] N. Ide and J. Pustejovsky. What does interoperability mean, anyway? Toward an operational definition of interoperability. In *Proc. Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong, China, 2010.

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

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 .
}
```

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