Undefined 1 (2009) 1–5
IOS Press

Linguistics Linked Data Life Cycle for Linguistics Resources Enhanced with Geospatial Data

Richard Littauer [a,b], Boris Villazon-Terrazas [c] and Steven Moran [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] *Intelligent Software Components, iSOCO, S.A., Av. del Partenon 16-18, Madrid, Spain*
E-mail: bvillazon@isoco.com
[d] *Research Unit Quantitative Language Comparison, Ludwig Maximilian University, Geschwister Scholl Platz 1, D-80539 Munich, Germany*
E-mail: bambooforest@gmail.com

Abstract. The Linguistics Linked Open Data cloud (LLOD), created and maintained by the Open Linguistics Working Group, is a (sub-)cloud of the LOD cloud, conforming to the Linked Open Data paradigm. The potential of a very large, interlinking, interoperable sub-cloud for linguistics research is great; however, early adopters may be hesitant to upload their datasets or use the cloud, due to a large learning curve or a lack of obvious uses. Here, we present an iterative and incremental Linguistics Linked Data Life Cycle that covers linguistic data resources, i.e., spreadsheets, that are enhanced with geospatial information. We use only freely accessible technologies in the Semantic Web framework, as well as a dataset of lexical and geospatial information of Dogon languages in West Africa. We also present a visualisation of language data from the World Atlas of Language Structures available as dataset within the LLOD. By doing so, we shed light upon the possibilities of the Semantic Web, and in particular the LLOD, for potential researchers in the digital humanities and computational sciences.

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

1. Introduction

The Semantic Web presents many opportunities for computational linguists and digital humanitarians interested in accessing, presenting, and discovering new information. However, the learning curve for the relevant technologies is steep, and first adopters may be hesitant due to lack of examples of possible uses and ignorance of useful databases for research. The Open Linguistics Working Group (OWLG) [3], an open community of researchers dedicated towards providing and disseminating openness in the broad field of linguistics, are creating and maintaining the Linguistics Linked Open Data (LLOD) [2,4] cloud for that purpose; to make clearly available and freely accessible multiple open databases for use by researchers and commercial companies alike.

The LLOD is not only freely accessible to researchers, but also openly adaptable for new databases and ontologies. All databases included in the LLOD must subscribe to the Linked Open Data paradigm [1], which demands that:

1. Referred entities should be designated by using Uniform Resource Identifiers (URIs),[1]
2. these URIs should be resolvable over HTTP,

[1] `http://tools.ietf.org/html/rfc3986`

3. data should be represented by means of specific W3C standards[2] (such as RDF), Štem 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, that they can be accessed and interpreted, and that entities that are associated on a conceptual level are also physically associated with each other.

Furthermore, 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. The ontology can be viewed in Fig. 1

In order to encourage future submission of resources into the LLOD, we present a new resource - a lexical and geospatial database of languages from the Dogon family, in Western Africa. According to the iterative and incremental Linked Data Life Cycle [11], all linked data follows a cycle: specification, modelling, generation, publication, and exploitation (which then feeds back into future specification). Here, we go through each of these stages on the way from taking the Dogon spreadsheet to an interactive visualisation using map4rdf. We also present a visualisation using a database already in the LLOD, the World Atlas of Language Structures (WALS) [6], by querying for geographic information for languages using a SPARQL endpoint for the LLOD (see Littauer *et al.*, this issue). We hope that these efforts will encourage future researchers to both add to and utilise the LLOD for their own research.

2. Related Work

Regarding language visualisation, there has been some work on displaying language differences on a broad scale. This includes work presenting hierarchical or cross-linguistic data [8,9], displaying related languages (gathered from WALS) by geographical proximity and relatedness [7], displaying word meaning with a world map [10], and displaying language locations on a globe [6]. However, the authors are not aware of any work using maps derived from data stored in RDF as here, although there have recently been more computational visualisations displaying language relatedness and dialectology using lexical items and location together [12].

3. LLD Life Cycle

In this section we present the specialisation of the Linked Data Life Cycle presented in [11] applied to linguistics resources enhanced to geospatial information.

3.1. Linguistics Resources

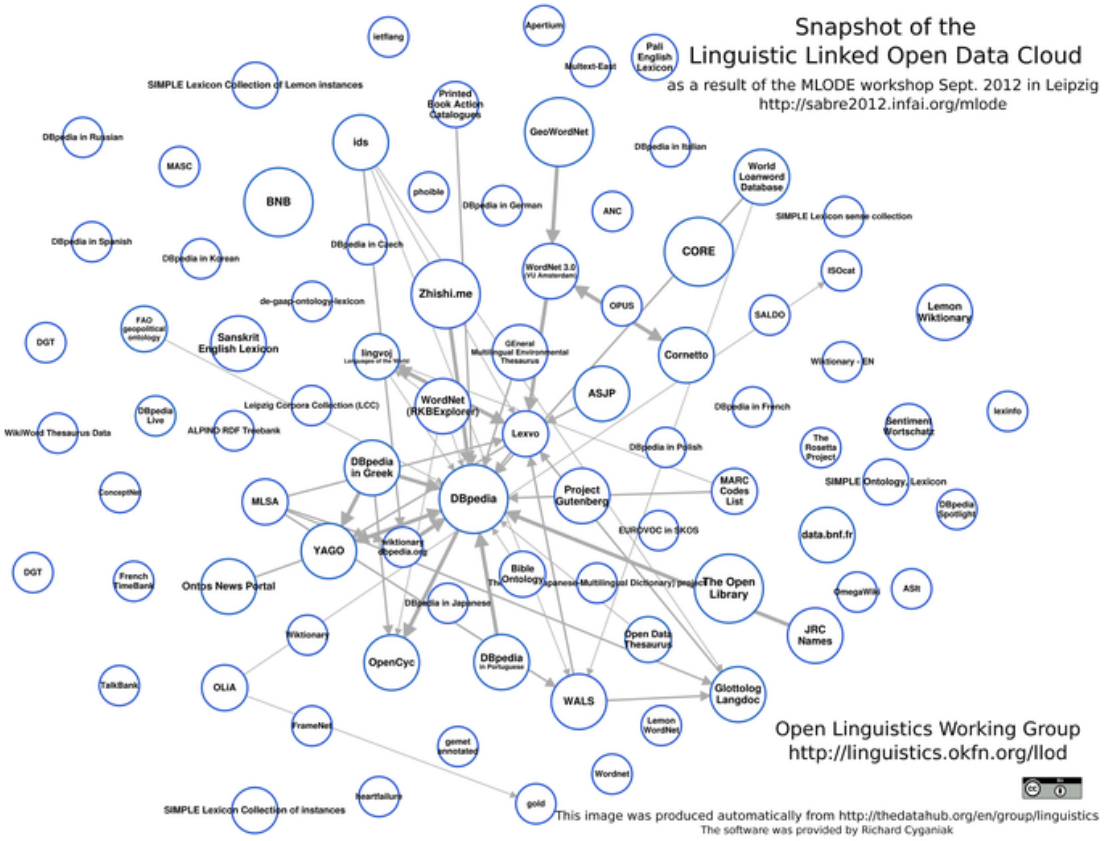
Our initial data source consisted of a spreadsheet containing GPS and lexical information for a variety of languages in the Dogon family, spoken in West Africa, and in particular in Mali. The information was gathered by linguists, and In particular, these fields were gathered for each datapoint: language family, family code, language group, language group code, language autonym, dialect code, ISO 639-3 code, country code, village name, major cite, population, latitude, longitude, social information, images, image descriptions, video, video transcription, and a comment on the language. This means that each data point had a particular GPS mapping, which corresponded roughly to a local village or town. The initial goal of this dataset was to understand the spread, diversity, and phylogenetic origins of Dogon languages. The classification of Dogon is unclear, although there is substantial amounts of work in this area.

3.2. Specification

In order to accurately model data according to the representations set forth by the communities involved in the Semantic Web, a specification must be laid out. A specification is a specific set of requirements that need to be satisfied by a resource. Here, we are concerned with converting plain text data from a spreadsheet into RDF. In order to be properly specified, we need to plan how the information will be represented. In order to fit the Linked Open Data paradigm de-

[2] http://www.w3.org/

Fig. 1. The LLOD Cloud



scribed in 1, all referred entities must be designated using a Uniform Resource Identifier (URI). Each URI needs to be resolvable over HTTP; that is, they need to be able to be looked up and referred to by a web page. This is the base URI structure. Vocabulary elements - what constitutes the language used to refer to different objects from this base URI - must also be defined, just as instances - particular cases of an element - need to be. The finished specification for our data set - which, at heart, is little more than the links needed to refer to the data once it is loaded into an ontology online, is described below.

3.2.1. URI design

All the resources in our dataset are defined using a URI. URIs have been designed with simplicity, stability and manageability in mind, following common guidelines for their effective use[3],[4], [5].

[3] `http://www.w3.org/TR/cooluris/`
[4] `http://www.w3.org/Provider/Style/URI`
[5] `http://www.w3.org/TR/chips/`

3.2.2. Base URI structure

`http://linguistic.linkeddata.es/`

3.2.3. Vocabulary elements

`http://linguistic.linkeddata.es/ontology/{property|class}`

For example:
`http://linguistic.linkeddata.es/ontology/officialName`

3.2.4. Instances

`http://linguistic.linkeddata.es/dataset/resource/{r.type|r.name}`

For example:
`http://linguistic.linkeddata.es/mlode/Village/Sokoura`

3.3. Modelling

The development of the linguistics vocabulary, which covers the linguistic information stored in the

resources described in section 3.1, was performed following an iterative approach based on the reuse of existing knowledge management resources for linguistics. In a nutshell we have reused the following vocabularies:

- GOLD[6], an ontology for descriptive linguistics.
- WGS84 Geo Basic Vocabulary[7], for representing the geospatial data.

Nevertheless, it was necessary to create some of our own terms. However, this work was largely trivial.

3.4. Generation

Next, the RDF was generated within three main phases. This is the process by which the spreadsheet turns into a useable resource. Here, we describe each one of these phases:

1. Importing the spreadsheet into MySQL database by using the MySQL importing tool.
2. Defining a set of R2RML[8] mappings among the MySQL database and RDF vocabulary elements.
3. Executing the R2RML engine, morph[9], for generating the RDF instances from the MySQL database data by using the defined R2RML mappings.

3.5. Publication

The generated RDF was then stored in Virtuoso[10] open source version. Virtuoso is essentially the server, allowing the RDF data loaded in the Generation step above to be queried using an endpoint. Virtuoso integrates with Pubby[11] to publish the results - Pubby is a fronted for endpoints, allowing users to query the data in the Virtuoso server using the SPARQL query language. Once Virtuoso and Pubby are running on data that has been loaded in using the R2RML engine, all of the data is essentially available for humans and computers to read. At this point the data, if it has been specified correctly, if the URIs are HTTP resolvable and if the vocabulary followed set standards, the process of lifting data into the Semantic Web is practically done. What's left is actually exploiting this data.

[6] http://linguistic-ontology.org
[7] http://www.w3.org/2003/01/geo/
[8] R2RML is a standard RDB2RDF mapping language http://www.w3.org/TR/r2rml/
[9] https://github.com/boricles/morph
[10] http://virtuoso.openlinksw.com/
[11] http://www4.wiwiss.fu-berlin.de/pubby/

3.6. Exploitation

The resultant dataset, following the previous steps, exposes the linguistics resources we first described enhanced with geospatial information, allowing for queries that otherwise require a lot of time by just looking at the original files. However, the data still hasn't been mapped - at best, SPARQL will return triples for any query.

However, by extending the queries presented previously, we have built an application[12] for showing each of the Dogon villages on the map by using the tool map4rdf[13] [5]. Figure 2 shows the application. At this point, each of the data points can be visualised.

4. Querying to Geospatial Information

We also used RDF data gathered automatically from the World Atlas of Language Structures (WALS) [6]. WALS can be freely accessed from http://wals.info, and the database can be downloaded as a .csv file. However, there is also an RDF endpoint, which has been integrated with the LLOD cloud. To access WALS data, then, we used the SPARQL endpoint available at: http://mlode-sparql.nlp2rdf.org/sparql.

In order to use the LLOD through 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.

Using this method, we devised a SPARQL query that gathered the longitude and latitude of each language within WALS specified by the three-letter language identifier WALS uses. Using the same methods as above, we converted this data into RDF, and then plotted it with the map4rdf software. The results of this can be seen here: http://geo.linkeddata.es/map4rdf-wals/#dashboard.

For more details on the SPARQL endpoint for the LLOD, see Littauer *et al.*, this issue.

[12] http://geo.linkeddata.es/map4rdf-dogon/
[13] https://github.com/boricles/linked-data-visualization-tools

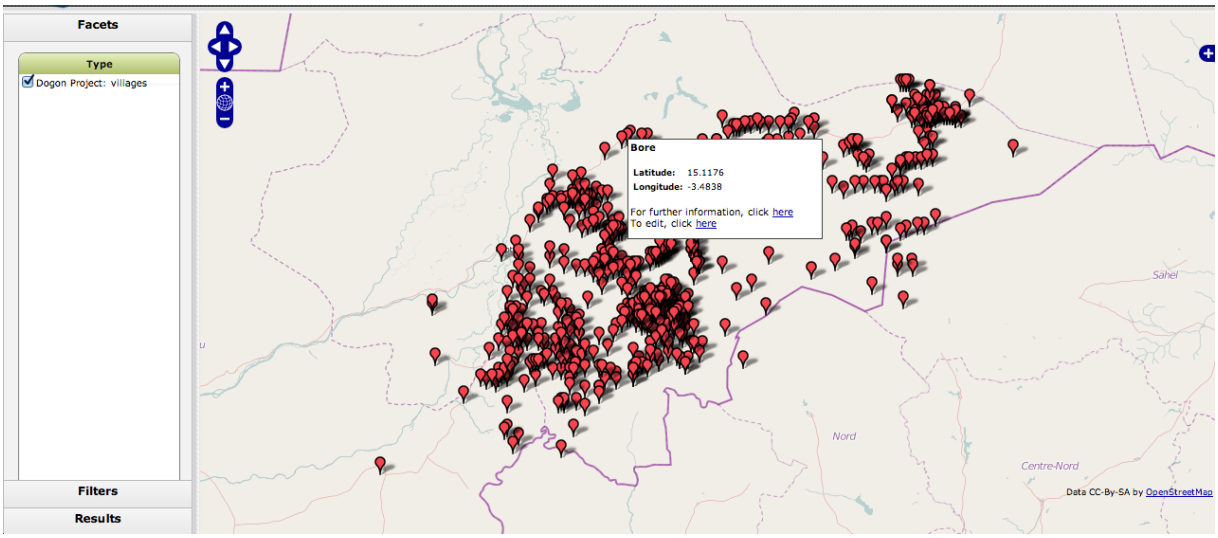


Fig. 2. Visualization of the Dogon villages

5. Conclusions

We have shown here the workflow necessary to upload data from a spreadsheet into an RDF database and then queried using a SPARQL endpoint, and an application on this data to show geospatial mapping. Any data uploaded using the process described up to Publishing 3.5 can be loaded into the LLOD cloud; the exact procedure for adding databases to the LLOD according to the specifications laid out in 1 can be found on the OWLG website[14]. The LLOD already contains dozens of individual databases which are interoperable and linked - and, as shown in 4, by using a SPARQL endpoint, one can immediately gather relevant data. It is hoped that the processes described here will be useful to future researchers interested in massive ontologies and the potential of the Semantic Web.

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] A. de León, F. Wisniewki, B. Villazón-Terrazas, and O. Corcho. Map4rdf - Faceted Browser for Geospatial Datasets. In *Proceedings of the First Workshop on USING OPEN DATA*. W3C, June 2012.

[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] R. Littauer, R. Turnbull, and A. Palmer. Visualising typological relationships: Plotting wals with heat maps. In *Proceedings of the EACL 2012 Workshop on the Visualization of Linguistic Patterns*, page 4, Avignon, France, April 2012. Association for Computational Linguistics.

[8] C. Rohrdantz, M. Hund, T. Mayer, B. Wälchli, and D. A. Keim. The world's languages explorer: Visual analysis of language features in genealogical and areal contexts. *Comp. Graph. Forum*, 31(3pt1):935–944, June 2012.

[9] C. Rohrdantz, T. Mayer, M. Butt, F. Plank, and D. A. Keim. Comparative visual analysis of cross-linguistic featuretures. In J. Kohlhammer and D. A. Keim, editors, *Proceedings of the International Symposium on Visual Analytics Science and Technology (EuroVAST 2010). The DEFINITIVE VERSION is available at diglib.eg.org.*, pages 27–32, 2010.

[10] R. Therón, L. Fontanillo, A. Esteban, and C. Seguín. Visual analytics: A novel approach in corpus linguistics and the nuevo diccionario histórico del español. In *III Congreso Internacional de Lingüística de Corpusi*, 2011.

[11] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological Guidelines for Publishing Government Linked Data Linking Government Data. In D. Wood, editor, *Linking Government Data*, chapter 2, pages 27–49. Springer New York, New York, NY, 2011.

[14] `http://linguistics.okfn.org/resources/llod/`

[12] M. Wieling, J. Nerbonne, and R. H. Baayen. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PLoS ONE*, 6(9):e23613, 09 2011.