Undefined 1 (2009) 1–5
IOS Press

Enhancing Linguistic Resources with Geospatial Mapping

Steven Moran [a], Richard Littauer [b,c] and Boris Villazon-Terrazas [d]

[a] *Research Unit Quantitative Language Comparison, Ludwig Maximilian University, Geschwister Scholl Platz 1, D-80539 Munich, Germany*
E-mail: bambooforest@gmail.com
[b] *Department of Intelligent Computer Systems, University of Malta, Msida, MSD2080, Malta*
[c] *Computational Linguistics Department, Saarland University, Saarbrücken, 66121, Germany*
E-mail: littauer@coli.uni-saarland.de
[d] *Intelligent Software Components, iSOCO, S.A., Av. del Partenon 16-18, Madrid, Spain*
E-mail: bvillazon@isoco.com

Abstract. The Linguistics Linked Open Data cloud, created and maintained by the Open Linguistics Working Group, is a (sub-)cloud of the Semantic Web, conforming to the Linked Open Data paradigm. The potential of a very large, interlinking, interoperable ontology for linguistics research is great; however, first 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 a research workflow, from a spreadsheet to RDF to visualisation, going through the entire iterative and incremental linked data cycle. 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 using an endpoint 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) [1], an open community of researchers dedicated towards providing and disseminating openness in the broad field of linguistics, have created and maintain the Linguistics Linked Open Data (LLOD) 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 [2], which demands that (1) Referred entities should be designated by using 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, 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.

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 cycle, 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 [3]. 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[4,5], displaying related languages (gathered from WALS) by geographical proximity and relatedness [6], displaying word meaning with a world map [7], and displaying language locations on a globe [3]. However, the authors are not aware of any work using maps derived from data stored in RDF as here, nor of visualisations aimed at displaying language relatedness using lexical items and location together.

3. Spreadsheet to geospatial information

Our initial dataset consisted of a spreadsheet containing GPS and lexical information for ....

This spreadsheet was then converted into data in RDF. We then used the tool map4rdf[1] to map it upon a globe, with an online portal at: `http://oegdev.dia.fi.upm.es/projects/map4rdf`. At this stage, we have plotted each of the Dogon villages on the map. Each point also contains additional information about the language spoken in that village. The map is available here: `http://geo.linkeddata.es/map4rdf-dogon/#dashboard`

4. Querying to Geospatial Information

We also used RDF data gathered automatically from the World Atlas of Language Structures (WALS) [3]. 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 the WALS, data, then, we used the SPARQL endpoint available at:

.`http://mlode-sparql.nlp2rdf.org/sparql`.

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.

We plotted this data with map4rdf software. The results of this can be seen here: `http://geo.linkeddata.es/map4rdf-wals/#dashboard`

5. Conclusions

References

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

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

[3] M. Haspelmath, M. Dryer, D. Gil, and B. Comrie, Eds., *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library, 2008. [Online]. Available: http://wals.info

[1] This can be accessed here: `http://code.google.com/p/map4rdf`

[4] 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*, vol. 31, no. 3pt1, pp. 935–944, Jun. 2012. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2012.03086.x

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

[6] 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*. Avignon, France: Association for Computational Linguistics, April 2012, p. 4.

[7] 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.