Linghub: Aggregated Metadata about Language Resources as Linked Data

John P. McCrae, Philipp Cimiano
CIT-EC, Bielefeld University
Bielefeld, Germany
{jmccrae, cimiano}@cit-ec.uni-bielefeld.de

Abstract

Abstract.

1 Introduction

Language resources are essential for nearly all tasks in natural language processing (NLP) and in particular for the adaptation of resources and methods to new domains and languages. In order to use language resources for new purposes they must first be discovered and this can only be done if there is a comprehensive list of all resources that may be available. To this there have been a number of projects that have attempted to collect such a catalogue using various methods and with differing degrees of data quality. We present a new portal, Linghub, that aims to integrate all these data from different sources by means of linked data and thus to create a portal, whereby all information about language resources can be included and queried using a common methodology. As such, this resource will enable wider discovery of language resources for researchers in NLP, computational linguistics and linguistics.

Currently, the approaches to metadata collection can be split into two broad classes: firstly, *curatorial* resources, which are those for which collections of language resources are maintained by one or more institute. Such resources have an advantage in that such metadata is normally of very high quality, however the resulting data often fails to cover the whole spectrum of data available. Examples of this include the META-SHARE (Federmann et al., 2012) project and the CLARIN project's Virtual Language Observatory (Van Uytvanck et al., 2012, VLO). On the other hand, *collaborative* approaches rely on data publishers self-reporting data about their own language resources. This can be advantageous as it allows reporting by researchers not directly collected to existing infrastructure projects, however the resulting data is often of lower quality as the systems may use free-text input or tagging input rather than controlled vocabularies, as they are easier for non-expert users to understand.

Given the nature of this difference we wish to make data available from multiple sources in a homogeneous manner and to this end we adopted a model based on the DCAT data model (Maali et al., 2014) along with properties from Dublin Core (Kunze and Baker, 1997). In addition, we used the RDF version (McCrae et al., 2015) of the META-SHARE model (Gavrilidou et al., 2012), to provide for metadata properties that are specific to language data and linguistic research. As such, in this paper we describe the creation of the largest collection of information about language resources and briefly describe its publication on the Web by means of linked data principles.

The rest of the paper is structured as follows...

2 Related Work

3 Extraction of data

4 Harmonization and duplicate detection

5 The Linghub portal

6 Conclusion

Acknowledgments

References

Christian Federmann, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli, and Marc Schröder. 2012. META-SHARE v2: An open network of repositories for language resources including data and tools. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation*, pages 3300–3303.

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Harris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz, et al. 2012. The META-SHARE metadata schema for the description of language resources. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation*, pages 1090–1097.

John Kunze and Thomas Baker. 1997. The Dublin Core metadata element set. RFC 5013, Internet Engineering Task Force.

Fadi Maali, John Erickson, and Phil Archer. 2014. Data catalog vocabulary (DCAT). W3C recommendation, The World Wide Web Consortium.

John P. McCrae, Penny Labrapoulou, Jorge Gracia, Marta Villegas, Víctor Rodríguez Doncel, and Philipp Cimiano. 2015. One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In *Proceedings of the 4th Workshop on the Multilingual Semantic Web*.

Dieter Van Uytvanck, Herman Stehouwer, and Lari Lampen. 2012. Semantic metadata mapping in practice: the virtual language observatory. In *LREC 2012: 8th International Conference on Language Resources and Evaluation*, pages 1029–1034. European Language Resources Association (ELRA).