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

There have been several attempts to collect metadata about language resources mostly associated with large infrastructure projects. CLARIN has been collecting resources under a project called the Virtual Language Observatory (Van Uytvanck et al., 2012), using the Component Metadata Infrastructure (Broeder et al., 2012, CMDI) to collect common metadata values from multiple sources. A similar project is META-SHARE (Piperidis, 2012) from the META-NET project where language resources are collected and high-quality, manual entries are created for each record. Similarly, the Open Languages Archives Community (Bird and Simons, 2003, OLAC) collects data from a number of sources although the metadata collected is not itself open. A

Source	Records	Triples
Datahub	185	10,739
LRE-Map	682	10,650
META-SHARE	2,442	464,572
CLARIN VLO	144,138	3,605,196
All	147,447	4,091,157

Table 1: Size of Linghub dataset by source

similar project called SHACHI has also collected some metadata (Tohyama et al., 2008). There has also been an attempt to track language resources by means of assigning them an International Standard Language Resource Number (ISLRN) similar to an ISBN used to track books (Choukri et al., 2012).

On the contrary some resources have instead collected data directly from creators of the resources, for example the LRE-Map (Calzolari et al., 2012) collects data from authors of papers submitted to conference, such as LREC. Similarly, Datahub [1] collects resources directly from those submitted to the website, but focusses primarily on linked data resources.

3 Extraction of data

In order to ensure that all the data from many sources can be queried in a homogenous manner we had to convert them to RDF. This process is also proved to be a valuable opportunity to align these vocabularies with standard vocabularies and fix any modelling errors. Two of our resources, LRE-Map and Datahub, were already available in RDF and thus, it should be the case that that the conversion of these resources required only renaming the URLs so that they would resolve without any collisions when uploaded to the Linghub portal. In fact, we also took this opportunity to fix a number of quality issues, such as fixing property values to either literals or URIs, reducing the number of blank nodes and changing modelling to that recommended in relevant standards, such as VOID (Alexander et al., 2011).

The other resources used XML schemas, for which we needed to create a custom conversion for each of them, which we did with the help of an invertible transformation language similar to XSLT. For META-SHARE, this was a challenging task as there were nearly a thousand unique tags defined and each one was examined to see if it was similar to an existing Semantic Web vocabulary, and in fact we ended up mapping to FOAF [2], SWRC [3] and the Media Ontology [4]. In the case of CLARIN, there was actually a significant difference between the XML schemas used by each contributing instance, with only a small common section giving the resource title and download link. We thus developed distinct mappings for the largest X institutes.

4 Harmonization and duplicate detection

Two key issues emerge when collecting data from a heterogenous set of sources such as we are doing. Firstly, the data is likely to be noisy and inconsistent in the properties it uses and more importantly in the values that these properties have. For example, languages may be represented by their English names or alternatively by means of the codes such as the ISO 639 codes. Secondly, it is often the case that a dataset may be recorded in multiple sources and thus, we may create multiple records of the same dataset. Furthermore, we often see duplication in the form of multiple records describing different sections of a single dataset or multiple usages of the single dataset. In order to remove these duplications we used state-of-the-art word sense disambiguation techniques, including Babelfy (Moro et al., 2014) to identify common controlled vocabularies and duplicate entries. For the case of properties we mapped to several existing resources, including LexVo (de Melo, 2013) for languages, and BabelNet for resource types. Duplicate entries were not removed from the dataset but instead were marked with the addition of the Dublin Core property *is replaced by*. In the case that these entries were subsets of resources the target of this link would be a new combined record for the entire resource and in the case of duplicate records collected from distinct sources we referred to the most complete triple, that is the record with the most triples.

5 The Linghub portal

In order to enable users to quickly and easily discover datasets, we set up a portal for browsing the dataset. Naturally we set this up as a site that publishes the individual records as either RDF or HTML, with the actual content de-

[1] http://datahub.io

[2] http://xmlns.com/foaf/spec/

[3] http://ontoware.org/swrc/

[4] http://www.w3.org/TR/mediaont-10/

Property	Target	Links
Language	LexVo	92,717
Type	BabelNet	12,139
Usage	META-SHARE OWL	1,310
Rights	License Ontology	103
Backlink	Original Resources	146,765

Table 2: Number of introduced links in the Linghub data

Figure 1: A screenshot of the Linghub interface

livered to the client decided by means of content negotiation. In addition, we provide a number of mechanisms by which users and automated agents can discover a dataset. In particular, for users we allowed browsing by means of faceted browsing of principle aspects of datasets including language. In addition, we enabled a full text search of the data based on the description. Machine based agents may access the endpoint by means of SPARQL querying, although the endpoint limits the agents to a very specific subset of the SPARQL query language. This is to ensure that the SPARQL querying remains stable and consistent, as full SPARQL queries could easily destabilize the server and adding timeouts could lead to unpredictable failures. In addition the server returns SPARQL JSON results and so can be called easily from a web browser.

6 Conclusion

Linghub is a new site that collects data from a large number of sources and makes it queriable through a common mechanisms. Furthermore, the data has not only been converted to RDF it has also been homogenized and linked to other bubbles in the Linguistic Linked Open Data Cloud. As such, this resource is likely to pay a pivotal role in enabling not only humans but also software agents to find new resources and use them for applications in natural language processing and artificial intelligence.

Acknowledgments

References

Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 2011. Describing linked datasets with the VoID vocabulary. Technical report, The World Wide Web Consortium. Interest Group Note.

Steven Bird and Gary Simons. 2003. Extending dublin core metadata to support the description and discovery of language resources. *Computers and the Humanities*, 37(4):375–388.

Daan Broeder, Menzo Windhouwer, Dieter Van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a component metadata infrastructure. In *Describing LRs with metadata: towards flexibility and interoperability in the documentation of LR workshop programme*, page 1.

Nicoletta Calzolari, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo, and Claudia Soria. 2012. The LRE Map. Harmonising community descriptions of resources. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation*, pages 1084–1089.

Khalid Choukri, Victoria Arranz, Olivier Hamon, and Jungyeul Park. 2012. Using the international standard language resource number: Practical and technical aspects. In *Proceedings of the 8th Interna-*

tional Conference on Language Resources and Evaluation, pages 50–54.

Gerard de Melo. 2013. Lexvo.org: Language-related information for the linguistic linked data cloud. *Semantic Web*, page 7.

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*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Stelios Piperidis. 2012. The META-SHARE language resources sharing infrastructure: Principles, challenges, solutions. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation*, pages 36–42.

Hitomi Tohyama, Shunsuke Kozawa, Kiyotaka Uchimoto, Shigeki Matsubara, and Hitoshi Isahara. 2008. Shachi: A large scale metadata database of language resources. In *Proceedings of the 1st International Conference on Global Interoperability for Language resources*, pages 205–212.

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).