Semantic Web 0 (0) 1
IOS Press

Multilingual Linked Data

John P. McCrae [a,*], Steven Moran [b], Sebastian Hellmann [c], Martin Brümmer [c]
[a] *CITEC, Bielefeld University, Germany. Email:* jmccrae@cit-ec.uni-bielefeld.de
[b] *University of Zurich, Switzerland. Email:* steven.moran@uzh.ch
[c] *AKSW Research Center, University of Lepizig, Germany. Email:*
(hellmann|bruemmer)@informatik.uni-leipzig.de

Abstract. The interaction of natural language processing and the Semantic Web have lead to the creation of a new paradigm known as Linguistic Linked Open Data (LLOD), whereby traditional language resources are made available as linked data. Conversely, the publication of corpora, machine-readable dictionaries as linked data has opened new resources to Semantic Web researchers and allowed new tools to be developed. In this special issue, we present recent development of tools and resources for creating and publishing language resources as linked data and tools to exploit this data to enable a multilingual Semantic Web.

Keywords: natural language processing, linked data, multilingualism, corpus, lexicon, linguistics

1. Introduction

In recent years, researchers in natural language processing (NLP) and linguistics have discovered Semantic Web technologies and employed them to better publish and connect their resources. It has been shown that linked data allows better data integration than existing models of linguistic data, due to the ecosystem of tools provided by the Semantic Web, such as queryand federation. In addition, the Semantic Web has already been used by several authorsto define data categories and enable better resource interoperability. The utility of this method of publishing language resources has lead to the interest of a significant sub-community in linguistics.

Language resources consists of a wide range of multilingual and monolingual data that may be of particular interest to researchers in linguistics or those building natural language processing systems. These resources can broadly be split into four main categories: Firstly, corpora which consist of collections of texts, with or without annotations or intra-document links; secondly, lexical resources describe the nature of single words, multiword expressions or terms; thirdly, language descriptions contain broad information about languages (e.g., grammars) as well as inter-language comparison. Finally, there is a group of resources, which capture metadata about language resources. It is important to note that all of the above resource types are generally in multiple languages and often feature multimedia content.

The papers in this special issue follow on from the 2012 Multilingual Linked Open Data for Enterprise (MLODE) Workshop which was held on the 25th September 2012 in Leipzig, Germany as part of the 2012 Software Agents and Services for Business, Research and E-Sciences (SABRE) conference. During the workshop, practical development and collaboration was established between many institutes that lead to papers presented in this issue. In addition, the workshop hosted the Monnet Challenge, a competition to develop the best new linked data language resourses, which was won by Gilles Sérraset of the DBnary project, whose resource is described in this special issue.

2. The Linguistic Linked Open Data Cloud

One of the major efforts to document and measure the growing adoption of linked data in computa-

*

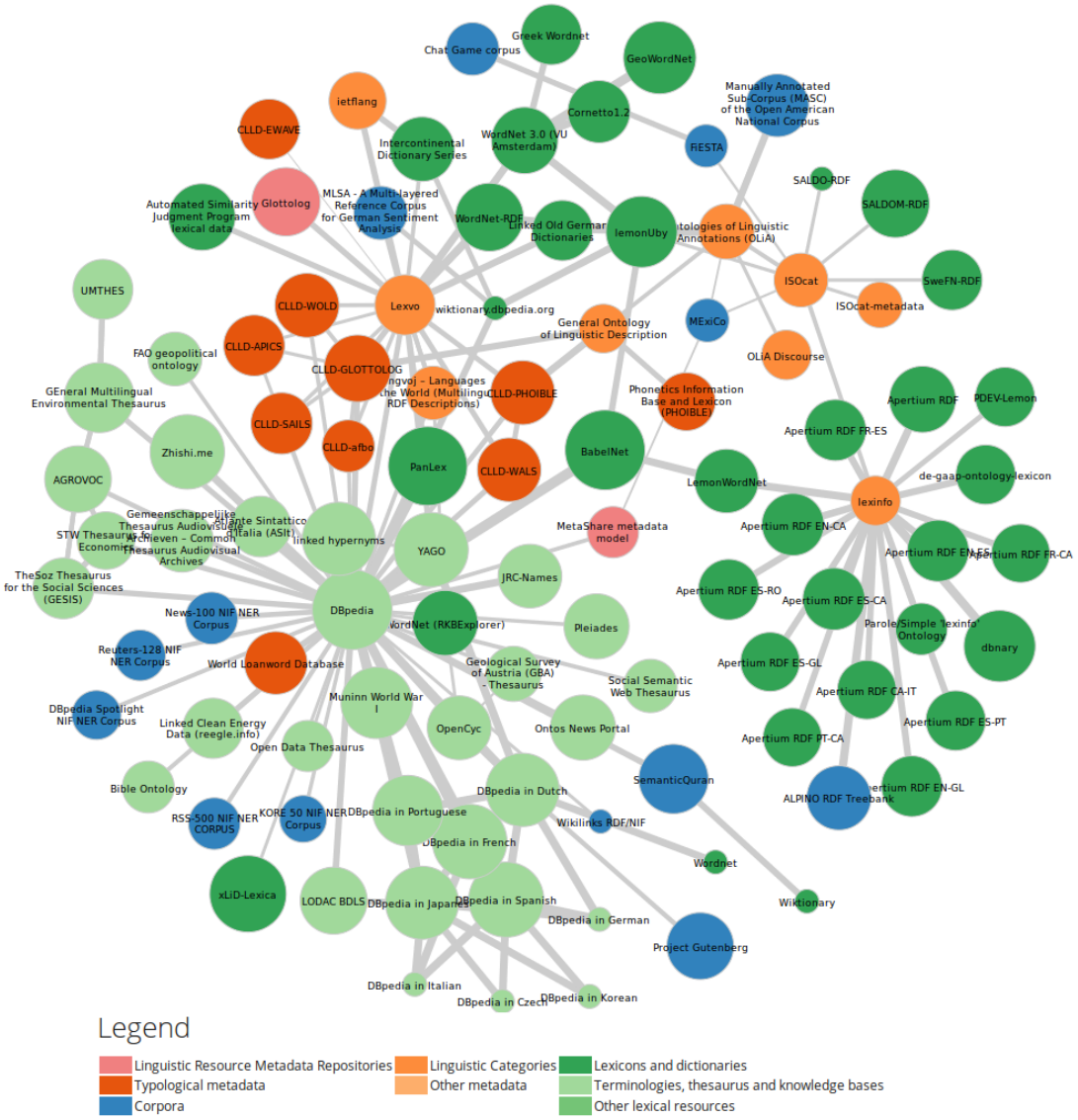

Fig. 1. The Linguistic Linked Open Data Cloud as of November 2014

tional linguistics has been by means of the Linguistic Linked Open Data (LLOD) Cloud Diagram (cf. Figure 1) for which the most recent version may be found at and `http://linguistic-lod.org`. This resource similar to the Linked Open Data (LOD) Cloud Diagram, depicts resources and the links in between them, however in this case the resources are limited to those of relevance to computational linguistics. In the LLOD diagram, we three authorities (i.e. nodes with a high number of inbound links) are evolving: in addtion to DBpedia –the existing nucleus of the main cloud– two resources, i.e. LexInfo and LexVo, provide important local nuclei for linguistic linked data. The former resource, is used by several papers in this special issue, and the second is described by its own paper in this special issue.

In this special issue we have collected a number of papers that describe the state of research into multilingual linked data and which exemplify the gamut of language resources. Firstly, Emilio Labra Gayo et al. describe practical design consideration of publishing multilingual linked data; this work is now continued by the W3C community group on Best Practices for Multilingual Linked Open Data. Next, Sherif and Ngonga

Ngomo describe their linked data representation of the Quran, a highly multilingual corpus. Several papers in this issue, namely Sérrasset, Villegas and Bel, Eckle-Kohler et al. and Del Gratta et al., describe the publishing of lexical resources drawn from existing projects such as PAROLE, Uby or Wiktionary. Then, Westphal et al. describe PanLex a resource, which describes over 7,000 languages and as such is vital for work in comparative linguistics and anthropology. Finally, de Melo and Chiarcos describe two distinct and complementary resources for describing languages and linguistic data.

As can be seen from this special issue, the growing adoption of linked data methods in linguistics has enabled new and interesting research paths not only for researchers in linguistics but also the Semantic Web. We believe that this continued collaboration between the communities can only lead to new and exciting discoveries and improved systems built on top of this data.

Acknowledgments. This special issue was supported by a grant from the EU's 7th Framework Programme provided for the project LIDER (GA no. 610782). We would like to thank all members of the Working Group for Open Linguistics (OWLG, `http://linguistics.okfn.org/`) and the DBpedia Community for their commitment to building the LOD Cloud.