Title: 3LD: Towards high quality, industry-ready Linguistic Linked Licensed Data

Summary. The application of Linked Data technology to the publication of linguistic data promises to facilitate interoperability of such resources and has lead to the emergence of the so called Linguistic Linked Data Cloud (LLD) in which linguistic data is published following the Linked Data principles. Three essential issues need to be addressed for such data to be easily exploitable by language technologies: i) appropriate machine-readable licensing information is needed for each dataset, ii) minimum quality standards for Linguistic Linked Data need to be defined, and iii) appropriate vocabularies for publishing Linguistic Linked Data resources are needed. We propose the notion of Licensed Linguistic Linked Data (3LD) in which different licensing models might co-exist, from totally open to more restrictive licenses through to completely closed datasets.

Type of contribution: Impact

Contributors' names and short Cvs.

Asunción Gómez-Pérez is Full Professor at UPM, director of the Artificial Intelligence department, director of the OEG and PhD in Computer Science (1993). Before joining UPM, she was visiting (1994-1995) the Knowledge Systems Laboratory at Stanford University. She also was the Executive Director (1995-1998) of the AI Laboratory at the School of Computer Science. She has coordinated SEALS, SemSorGrid4Env and Ontogrid and she has participated in more than 15 EU projects. Her main research interests are ontologies and the semantic Web. She is the coordinator of the Lider project.

Daniel Vila-Suero is a PhD candidate at the Ontology Engineering Group (Universidad Politécnica de Madrid, Spain), and MSc in Computer Science. His research topics are multilingualism in the Web of Data, digital libraries, methodologies and Linked Data. He participated in the Spanish research projects related to Linked Data and multilingualism datos.bne.es and BabeLData as well as in several standardization groups of the W3C and IFLA.

Víctor Rodríguez-Doncel is a researcher in the Ontology Engineering Group (Universidad Politécnica de Madrid, Spain). He received his PhD in 2010 from the Universitat Politècnica de Catalunya, having conducted research in the academy (Aristotle University of Thessaloniki in Greece, Universitat Pompeu Fabra) as well as in the industry (GMV, Barcelona Digital). He is

editor of the ISO/IEC 21000-19 international standard, the Media Value Chain Ontology for representing the intellectual property along the multimedia value chain. He is also author of two books, several JCR indexed journal papers and has participated in FP6 and FP7 Framework projects (VISNET-I, AXMEDIS, VISNET-II, SUPERHUB).

John McCrae received a MSci in Mathematics and Computer Science from Imperial College London, UK in 2006 and a PhD from the National Institute of Informatics, Japan in 2009. In 2009 he joined the Semantic Computing Group at CITEC in the University of Bielefeld, where has worked on the Monnet project and is now working on the LIDER project.

Philipp Cimiano is full professor for computer science at the University of Bielefeld. He leads the Semantic Computing Group and is affiliated to the Center of Excellence on "Cognitive Interaction Technology" (CITEC). He studied computer science in Stuttgart and received his doctoral degree and habilitation from the University of Karlsruhe. He is currently leading as co-chair the activities of the Ontology-Lexicon W3C Community Group.

Guadalupe Aguado de Cea is Associate Professor at UPM, MSc in Translation and PhD in English Philology. Her current research areas are Terminology, Translation, and ontologies. She is the Chair of the AENOR CTN_191 Terminology Committee, the Spanish Committee of ISO TC 37.

3LD: Towards high quality, industry-ready Linguistic Linked Licensed Data

Daniel Vila-Suero[1], Victor Rodríguez-Doncel[1], Asunción Gómez-Pérez[1], Philipp Cimiano[2], John P. McCrae, and Guadalupe Aguado-de-Cea[1]

[1] Ontology Engineering Group, Facultad de Informática, UPM. Madrid, Spain
{dvila, vrodriguez, asun, lupe}@fi.upm.es
[1] Forschungsbau Intelligente Systeme (FBIIS). Universität Bielefeld. Bielefeld, Germany
{cimiano, jmccrae}@cit-ec.uni-bielefeld.de

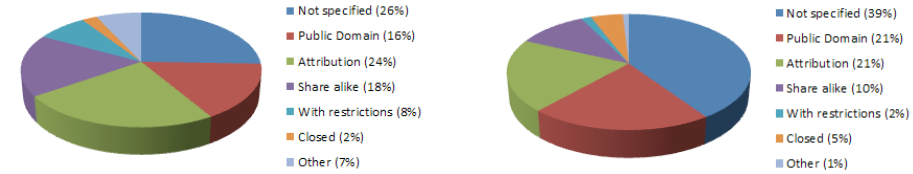
Abstract. The application of Linked Data technology to the publication of linguistic data promises to facilitate interoperability of such resources and has lead to the emergence of the so called Linguistic Linked Data (LLD) Cloud in which linguistic data is published following the Linked Data principles. Three essential issues need to be addressed for such data to be easily exploitable by language technologies: i) appropriate machine-readable licensing information is needed for each dataset, ii) minimum quality standards for LLD need to be defined, and iii) appropriate vocabularies for publishing LLD resources are needed. We propose the notion of Linguistic Linked Licensed Data (3LD) in which different licensing models might co-exist, from totally open to more restrictive licenses through to completely closed datasets.

Keywords: Linked Data, Language resources, licenses, intellectual property rights, vocabularies.

1 Introduction

The last few years have witnessed an explosive growth in the amount of multilingual and cross-media digital contents (both structured and unstructured) available on the Web. The issue of managing this content effectively and extracting maximum value remains a major challenge for companies and researchers. In this context, the Linked Data (LD) paradigm offers new mechanisms for organizations to share, connect and exploit data more efficiently on the Web. LD is a set of practices and standards for publishing and consuming structured data on the Web which focuses on identifying data items with HTTP URIs, describing them in machine-readable formats such as RDF, and linking them to other data items using hyperlinks and creating a distributed data network.

Data published as LD can be licensed in different forms. Some of the LD has been released as Linked Open Data, referring to the fact that it is open – being either in the public domain or licensed under relatively open terms, e.g. under creative commons share-alike licenses. Some institutions also publish data with more restrictive licenses or even in a completely closed fashion. The study presented in [1] analyzed 1836 linked datasets registered in *datahub.io* in May 2013, 338 of them being labeled as 'LOD datasets'. The results, shown in Figure 1, reveal that a significant number of datasets had been published without a license or under restrictive terms.



(a) Licenses in datasets present in datahub.io (b) Licenses in datasets present in the LOD cloud

Fig. 1. Kind of licenses used in (a) datasets present in datahub.io and (b) those annotated present in the LOD cloud as of May 2013 [2].

Researchers and organizations working on linguistic resources have also shown great interest in publishing their data as LD to facilitate the sharing and integration of various datasets. Thus, the term Linguistic LOD (LLOD) has been coined to refer to a new LOD-based ecosystem of free, interlinked, and semantically interoperable linguistic resources [2]. This collection of resources, collected by the "Open Data in Linguistics" group[1] contains 86 resources which have been analyzed in this work:

(1) The resources *include annotated linguistic corpora, lexical databases and lexical-semantic resources*, but also *others only tangentially relevant* for the linguistic community like Dbpedia or data from other domains (libraries) ; (2) The datasets included in the LLOD *are made available in different manners*. In particular, 62 datasets provide some sort of access to RDF data (via a SPARQL endpoint, URIs of RDF resources, a VoID file or RDF data dumps), while the other 24 dataset's descriptions do not provide a clear entry point to RDF data ; (3) The *vocabulary used to describe the linguistic resources is not uniform*. For instance, the *LemonWordNet* uses Lemon and the *RKBExplorer WordNet* uses the Wordnet 2.0 RDF/OWL representation[2], and ; (4) The *licenses used* in the 86 resources *are distributed* as follows: 34 resources are in the public domain or require attribution, 27 of them are share-alike (requiring derived resources to stay under the same terms) and 25 of them being given under more restrictive terms, with a missing license or directly closed. Table 1 shows a summary where similar licenses have been grouped in 7 categories of increasing restrictiveness, and not all resources included in the LLOD are open.

Table 1. Kind of licenses used in the Linguistic Linked Open Data cloud

Kind of License	**License types**	**Number (86)**
Public Domain	other-pd,cc-zero,odc pddl	11
Attribution	cc-by, odc-by, other-at	23
Share Alike	cc-by-sa, gfdl, odc-odbl	27
Non-Commercial / Non Derivative	cc-nc, other-nc, gpl-2.0	6
Other open	other-open	8
Not specified	notspecified	5
Closed	other-closed	6

Since the Linguistic LOD as documented by the OKFN is at an early stage, there are initiatives like the one proposed in the LIDER project[3] that aims at supporting and complementing this initiative by developing guidelines and best practices that help publishing linguistic resources as LD. These guidelines and best practices are aimed

1 http://linguistics.okfn.org/
2 http://www.w3.org/TR/wordnet-rdf/
3 http://lider-project.eu

at providing advice and mechanisms to facilitate the exploitation of linguistic resources by third party applications. In this paper we discuss three of the most pressing needs in order to realize this vision, the need: i) for licensing models and mechanisms to encourage data exploitation (see Section 2), iii) of minimum quality standards for Linguistic LD (see Section 2), and iii) the need of using open, standard and high quality data models for representing linguistic data (see Section 3).

2 3LD: Linguistic Linked Licensed Data

The main idea of LD is to create an ecosystem that facilitates the browsing, discovery and exploitation/reuse of datasets for applications. The issue of whether datasets can be actually used for a specific purpose is thus a crucial one, so that understanding the conditions under which a certain dataset has been licensed is crucial. Such licensing information should be expressed ideally in a machine-readable fashion to facilitate automated reasoning by end applications on the conditions of use of a dataset.

The second important aspect from the perspective of reuse is quality. Clearly, end users would only want to use datasets that have a minimum level of quality, use of standard vocabularies and links to other relevant resources among other features.

Focusing on these two issues, we will refer to 3LD as the set of linguistic resources expressed as LD that have a minimum quality and are published along with a machine-readable, automatically discoverable license. The proposed best practices for declaring the license in a machine-readable, standard way include:
(1) Adding the proper rights metadata, either in a separated DCAT or VOID file or within the resource (for the case of ontologies).
(2) Using standard predicates to declare the rights information of a resource, specifically dct:license to specify a license (using the elements of the Dublin Core vocabulary[4]) and dct:rights to inform about additional rights information.
(3) Using standard licenses, like those published by the Open Data Commons (ODC) or Creative Commons (CC), referencing them by their URI.
(4) Using ad-hoc rights expressions when the CC/ODC licenses do not suffice, specifically for all the non-open cases. These expressions should be made with rights expression languages, like the ODRL 2.0[5] or the LDR[6] (see Figure 2).

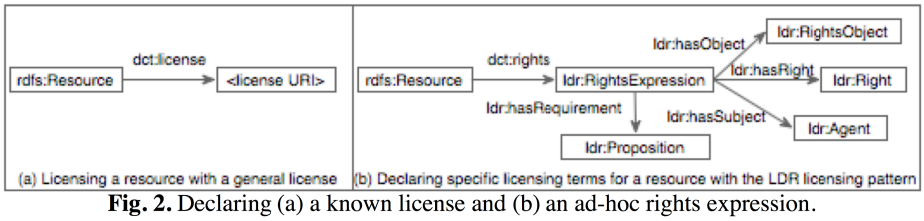


Fig. 2. Declaring (a) a known license and (b) an ad-hoc rights expression.

3 Data models for representing linguistic data on the Web

For publishing data as open LD, there is a clear need for the creation of agreed-upon / standard models for the representation of such data. Unfortunately, most of the existing models for the representation of linguistic data are not easy to adapt to the case of LD and are often not open standards. For example, the Lexical Markup Framework (LMF) [4] is a standard that would not be suitable for the representation of LD, for several reasons: firstly, it is not an open standard but requires payment to

4 http://purl.org/dc/terms/
5 http://www.w3.org/ns/odrl/2/
6 http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/

view even the specification documents. Secondly, the model is a meta-model meaning that data providers do not have any normative model to conform to, although an XML DTD-Schema is provided, which limits the representation of a lexicon to a single XML File. This is not suitable for the web where individual entries in a lexicon are likely to be represented on a single page. Finally, LMF does not support the use of URIs to identify concepts or interact well with other web standards such as OWL.

To address these issues, the OntoLex community group has been launched in 2012 to develop a new model based on lemon (Lexicon Model for Ontologies) [5], to represent lexical resources relative to ontologies. This model has been developed around the principles of being open, RDF-native, minimalist and linguistically sound. Similarly, the NIF (NLP Interchange Format) ontology [6] was developed as a model for representing stand-off annotation of corpora using RDF. These models address some forms of linguistic LD however do not cover all possible forms of linguistic data, and as such there is still a need to develop further models for multimodal resources, typological databases and many other kinds of linguistic data.

4 Conclusion

The application of LD technology to the publication of linguistic data promises to alleviate issues related to the integration and aggregation of dataset stemming from heterogeneous sources and using different vocabularies. This has lead to coning the term Linguistic LD for all linguistic datasets that are published following the LD principles. However, three essential issues need to be addressed for such data to be easily exploitable by language technologies. First of all, datasets need to be enriched with machine-readable licensing information so that applications can reason about conditions under which it is legitimate to use a particular resource. While in the general case open licenses are preferable and compatible with the Web-style publishing used in LD, some use cases and datasets might require more restricted licenses or even datasets to be closed while being at the same time linked. Second, linked linguistic datasets need to have a minimum quality in order to build trust by end applications. Finally, we need shared and agree-upon vocabularies and guidelines to foster standardization and thus easier exploitation of resources. We have briefly discussed these issues in this article and presented some preliminary investigations on the distribution of licenses in the Linguistic LD Cloud. Overall, we have coined the notion of a Licensed Linguistic LD in which different licensing models can co-exist, from totally open to more restrictive licenses through to completely closed datasets.

References

1. Rodriguez-Doncel, Victor, Gómez-Pérez, A. and Mihindukulasooriya, Nandana. Rights declaration in Linked Data. in Proc. of the 3rd Int. W. on Consuming Linked Data O. Hartig et al. (Eds) CEUR vol. 1034 (2013)
2. Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff. 2012. Linking linguistic resources: Examples from the Open Linguistics Working Group, In: Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), Linked Data in Linguistics. Representing Language Data and Metadata, Springer, Heidelberg, p. 201-216
3. Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. "Lexical markup framework (LMF)." In International Conference on Language Resources and Evaluation-LREC 2006. 2006.
4. McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia et al. "Interchanging lexical resources on the semantic web." Language Resources and Evaluation 46, no. 4 (2012): 701-719.
5. Hellmann, Sebastian, Jens Lehmann, and Sören Auer. "Linked-data aware uri schemes for referencing text fragments." In Knowledge Engineering and Knowledge Management, pp. 175-184. Springer Berlin Heidelberg, 2012.