DOI: https://doi.org//10.21747/16466195/ling16r2

Philipp Cimiano; Christian Chiarcos; John P. McCrae; Jorge Gracia (2020). *Linguistic Linked Data. Representation, Generation and Applications*. Springer International Publishing. ISBN 978-3-030-30225-2

Purificação Silvano
msilvano@letras.up.pt
Centro de Linguística da Universidade do Porto, Faculdade de Letras da Universidade do Porto (Portugal)

1. Background

During the last decades, there has been a proliferation of language resources, which are of utmost relevance to research in many areas from Linguistics to Natural Language Processing (NLP). However, frequently these resources lack structural and conceptual interoperability (Chiarcos *et al*. 2020), which are key conditions warranted in a Semantic Web where multilingual information from different resources should be more easily accessed, identified, and interpreted. Linguistic Linked Open Data (LLOD) enables such a reality.

The LLOD Infrastructure, envisioned by a working group of Open Language Foundation, the Open Linguistics Working Group (OWLG) (Chiarcos *et al*. 2011), combines two prior existing concepts, open data and linked data, and applies them to language resources. As other Web resources, linguistic resources must comply with the following Linked Open Data's directrices regarding publication and representation: (a) data openly licensed; (b) entities referred by URIs (Uniform Resource Identifier); (c) URIs resolvable over HTTP (Hypertext Transfer Protocol); (d) data represented by web standards, such as HTML (Hypertext Markup Language), RDF (Resource Description Framework) or JSON-LD (Javascript Object Notation); and (e) resources linked to each other (Chiarcos *et al*. 2013). These Linked Data Principles (Bizer *et al*. 2009) ensure homogeneity of the linguistic resources, and thus, a more global access to them, which, in turn, can lead to more

cooperation across disciplines (Chiarcos *et al*. 2020). In fact, in this manner, LLOD provides several benefits such as representation, by means of linked graphs, which allow a more adaptable representation format for linguistic data; interoperability; expressivity, due to vocabularies like Web Ontology Language (OWL) and Lemon; and dynamicity, as a result of the perpetual enhancement of web data (Chiarcos *et al*. 2013; https://linguistic-lod.org).

Throughout the years, the LLOD infrastructure has been enriched with contributions from several projects. *Creating knowledge out of Interlinked Data* (LOD2) was one of the first (2010-2014) to contribute to the development of LLOD infrastructure. More recently, *Ready-to-use Multilingual Linked Language Data for Knowledge Services across Sectors* (Prêt-à-LLOD), a European H2020 project, has been working towards further expansion of LLOD infrastructure, and its sustainability. Towards these goals, Prêt-à-LLOD is developing methods to discover, transform and link linguistic data to be published as LLOD, and is demonstrating the usability and efficacy of those methods in real-world problems of language technology industries (e.g. Semantic Web Company, Oxford University Press) by means of use case pilots (Declerck *et al*. 2020). Prêt-à-LLOD has also been collaborating closely with *European Lexicographic Infrastructure* (ELEXIS), which is creating a dictionary matrix based on LLOD. The COST action *European Network for Web-Centred Linguistic Data Science* (NexusLinguarum) aims to promote the construction of a global ecosystem of multilingual and semantically interoperable linguistic data by developing linked data technologies, combined with NLP techniques and multilingual resources (https://nexuslinguarum.eu/the-action), which will be of great value to the LLOD Infrastructure.

The LLOD Infrastructure has grown so significantly that, first, in 2014, the category "linguistics" was integrated in the Linked Open Data cloud, and second, in 2018, the LLOD cloud was incorporated as a domain-specific addendum (https://linguistic-lod.org). In fact, the LLOD cloud's increase has been higher than the LOD cloud (Chiarcos *et al*. 2020). At the moment of the writing, the language resources available in the LLOD cloud are 220, distributed in the following categories: corpora, lexicons and dictionaries, terminologies, thesauri and knowledge bases, linguistic resource metadata, linguistic data categories, typological databases and others.

The appeal of LLOD is substantiated by the increasing community of researchers working not only towards the creation or conversion of digital linguistic resources into linked data (LD), but also towards the development of standards and models to do so. Moreover, LLOD has stirred up the interest of researchers who are eager to learn more about this paradigm. It is in this regards that *Linguistic Linked Data. Representation, Generation and Applications*, by Philipp Cimiano, Christian Chiarcos, John P. McCrae and Jorge Gracia, who have authored seminal works in LLOD, is a must-read book.

2. The book

This book offers a thorough description of LLOD to students, teachers and researchers from Linguistics and Computational Sciences who work with language resources and are keen on learning about guiding principles and techniques to create, reconvert and publish those resources as linked data, and about possible applications of LLD resources. The book summarizes a wide array of contents about different aspects of LLD with clear explanations, illustrative examples, and references for additional information.

The book is very well organized in five parts: *Preliminaries*, *Modelling*, *Generation and Exploitation*, *Use Cases* and *Conclusions*. The first part is key to understand the rest of the book, as stated by the authors, because it presents the basic concepts underlying LLD. The second and third parts explain how to model language resources into LD and how to represent and to discover linking between datasets. The fourth part is dedicated to the description of different applications of LLD. The fifth part encompasses the conclusions. As one can infer from this overview, the book is quite overarching covering the necessary theoretical principles to comprehend the subject, but also mentioning use cases to demonstrate the benefits of adhering to the LLOD model.

Zooming in the structure of these five parts, each part comprises a number of chapters, and each chapter is divided into sections preceded by an *Abstract* and followed by *Summary and Further Reading*, and *References*. The abstract it very helpful, allowing the reader to gather more information

about the chapter's content before reading it. The summary gives the reader a systematization of the main points discussed in the chapter, which is always useful to consolidate what the reader learnt, while the further reading part specifies the references to broaden the reader's knowledge about the issues explored in the chapter.

The first three chapters integrate part I – *Preliminaries*. Chapter 1 demonstrates how LD is the perfect scenario to attain the *FAIR Principles*, that are, *Findability*, *Accessibility*, *Interoperability* and *Reusability*. In fact, the already mentioned Linked Open Data's directrices guarantee that language resources are findable, accessible, interoperable, and reusable. Chapter 2 focuses on RDF, the quintessentially format for LLD, describing its semantics and its most well-known serializations (e.g. Turtle (Terse RDF Triple Language), JSON-LD). The contribution of OWL to represent ontological knowledge, and the query language SPARQL (Protocol and RDF Query Language) are also explained. It goes without saying that the foundational nature of this chapter means that if the reader's intention is to model language resources resorting to RDF or Turtle, work with OWL or carry out queries using SPARQL, further study and practice are needed. The last chapter of the first part provides the motivation for LLD and outlines its beginning, the story behind LLOD cloud, and the substantial work of the LLOD community, namely of OWLG.

Chapters 4 to 8 compose part II – *Modelling*. The first chapter of this part gives a detailed account of the Ontolex-lemon, a generic lexicon model, which was developed by the very active Ontology-Lexicon Community Group, and which enables the representation of lexical information as linked data. Its popularity is proven by the large collection of dictionaries that use the lemon model (cf. DBpedia, Apertium, among others). Chapter 5 instructs the reader on how to annotate data according to RDF principles. It starts with a summing up of annotation languages used by NLP and Digital Humanities, such as the CoNLL format and TEI (Text Encoding Initiative), showing how these can be adapted to create LLD. The following section proceeds to discussing two compatible RDF-based formalisms for referencing data on the web: Web Annotation, and a simpler alternative, that is NIF (NLP Interchange Format). Chapter 6 complements the previous chapter by elaborating on the necessary vocabularies to represent linguistic

annotations in an interoperable manner, namely those grounded on OWL and RDF models, such as CoNLL-RFD and POWLA. The technologies and techniques to query the resulting annotated corpora are also revised. Since it is important to encode not only linguistic data, but also the information about the language resources that contain those linguistic data, chapter 7 refers the models for the representation of language resources metadata, so that they can be easily found and reused, two essential values of LLOD. It is the case of DC-Terms (DCMI Metadata Terms vocabulary) for general metadata (title, license, description creator…) and Meta-Share.owl ontology for information pertaining to the language resources (resource type, modality, number of languages…). The last chapter of the second part focuses on strategies to achieve conceptual interoperability, being one of those the laborious undertaking of harmonizing linguistic categories through the creation of general terminology repositories like ISOcat, and more specific ones, such as OLiA (Ontologies of Linguistic Annotation).

The next group of chapters (9-11) constitute part III – *Generation and Exploitation*. Chapter 9 takes the reader through the process of publishing LLD, itemizing and illustrating each of the six steps: (i) specification – analysis and description of data sources; (ii) modelling – creation or selection of vocabularies to represent in RDF; (iii) generation – production of RDF datasets from the data sources; (iv) linking – connection of RDF datasets; (v) publication – release of the datasets; (vi) exploitation – development of applications that use the datasets. Being at the core of LLD paradigm the linking concept, the techniques to represent the links between datasets in the same language and across languages, and to discover those links are, obviously, quite pertinent. For this reason, chapter 10 is fully devoted to link representation and discovery. Chapter 11 argues in favor of using the models compliant with LLD formalisms, like NIF, for creating NLP pipelines, and it also exemplifies how this workflow can be operationalized.

Finally, chapters 12-14 form part IV – *Use Cases*. The linking of multilingual wordnets is addressed in chapter 12, while the following chapter highlights the prominence of LOD in Digital Humanities. In both cases, the reader becomes acquainted with the broad range of applications that integrate LLOD technologies, in particular in the field of Digital Humanities. In chapter 14, the authors underline the hurdles of finding language

resources in the repositories, and of standardizing metadata pertaining to those repositories, and present Linghub, a portal that gathers metadata from different repositories, as a solution to overcome those obstacles.

Chapter 15 (Part IV) closes this incursion into the world of LLOD with a succinct recapitulation of the topics tackled in the book and with a glimpse into the required future steps towards the expansion of the LLOD infrastructure.

As it can be assumed from this brief synopsis, the book under review supplies the reader with an extensive survey about the topic of LLOD, divulging the models and best practices so that more researchers adhere to this paradigm, and together continue to invest in the development of the LLOD infrastructure.

REFERENCES

Bizer, C., Heath, T. & Berners-Lee, T. 2009. Linked data – The story so far. *International Journal on Semantic Web and Information Systems*. 5: 1-22.

Chiarcos, C., Hellmann, S. & Nordhoff, S. 2011. Towards a Linguistic Linked Open Data cloud: The Open Linguistics Working Group. *TAL (Traitement Automatique des Langues)*. 52(3): 245-275. Retrieved January 12, 2021, from the World Wide Web: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9AAA230DDF3A6AA313FE939EF95C30AC?doi=10.1.1.377.2076&rep=rep1&type=pdf

Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. 2013. Towards open data for linguistics: Linguistic linked data. In: Alessandro Oltramari, Piek Vossen, Lu Qin & Eduard Hovy (Eds.). *New Trends of Research in Ontologies and Lexical Resources: Ideas, Projects, Systems*. Springer, Berlin, Heidelberg, 7-25. Retrieved January 12, 2021, from the World Wide Web: https://link.springer.com/content/pdf/10.1007%2F978-3-642-31782-8.pdf

Chiarcos, C., Klimek, B., Fäth, C., Declerck, T. & McCrae, J. P. 2020. On the Linguistic Linked Open Data Infrastructure. In: Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajič, Stelios Piperidis & Andrejs Vasiļjevs (Eds.). *Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)*. Marseille, France: European Language Resources Association, 8-15. Retrieved January 12, 2021, from the World Wide Web: https://www.aclweb.org/anthology/2020.iwltp-1.2/

Declerck, T., McCrae, J. P., Hartung, M., Gracia, J., Chiarcos, C., Montiel-Ponsoda, E., Cimiano, P., Revenko, A., Sauri, R., Lee, D., Racioppa, S., Nasir, J., Orlikowski, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M., Khvalchik, M., Gonzalez, M. & Cooney, K. 2020. Recent developments for the Linguistic Linked Open Data Structure. In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Eds.). *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France: European Language Resources Association, 5660-5667. Retrieved January 12, 2021, from the World Wide Web: https://aclanthology.org/2020.lrec-1.695.pdf