The Linguistic Linked Open Data Cloud: Phenomenal Cosmic Powers... Itty Bitty Quality Space!

Pasquale Esposito[1,†], Maria Angela Pellegrino[1,*,†], Vittorio Scarano[1] and Gabriele Tuozzo[1]

[1]Università degli Studi di Salerno, via Giovanni Paolo II, 132, 84084 Fisciano (SA), Italy

Abstract

The Linguistic Linked Open Data movement aims to model linguistic data according to the Semantic Web technologies, exploiting interlinking, open licenses, and enabling accessibility. But as *every rose has its thorns*, Linguistic Linked Open Data are not immune to data quality issues. This **poster** paper aims to document the current status of the Linguistic Linked Open Data Cloud in terms of amount of data, licensing, and accessibility, to identify potentialities and limitations that limit its utility and exploitation.

Keywords

Linguistic Linked Open Data, Quality Assessment, Accessibility, Contextual dimensions

1. Introduction

In the rapidly evolving landscape of linguistic research, the advent of Linguistic Linked Open Data (LLOD) [1] has catalyzed a paradigm shift, heralding an era of unprecedented collaboration and knowledge exchange among linguists and the Semantic Web community. Spearheaded by pioneering efforts such as the Open Linguistics Working Group [2, 3] and initiatives like LingHub [4], researchers are harnessing the power of linked data to create a vast interconnected network of linguistic resources. Furthermore, LLOD have been successfully exploited in Natural Language Processing tasks, demonstrating the potential of LLOD to drive innovations at the intersection of linguistics and web technologies [5, 6].

However, this endeavor is not without its challenges. The LLOD ecosystem is characterized by a rich tapestry of resources, ranging from traditional linguistic databases to encyclopedic knowledge bases like DBpedia and Wikidata, leading to heterogeneity in both content and structure [7]. Additionally, the accessibility of some resources remains a hurdle, with certain datasets being unavailable or inadequately represented. As a result, estimating and monitoring the quality of LLOD is crucial. This **poster** paper aims to document the quality assessment of the LLOD Cloud in the direction of identifying potentialities and directions for improvement.

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

*Corresponding author.

†These authors contributed equally.

pasesposito@unisa.it (P. Esposito); mapellegrino@unisa.it (M. A. Pellegrino); vitsca@unisa.it (V. Scarano); gtuozzo@unisa.it (G. Tuozzo)

0009-0006-3464-5861 (P. Esposito); 0000-0001-8927-5833 (M. A. Pellegrino); 0000-0001-8437-5253 (V. Scarano)

2. Linguistic Linked Open Data Quality Assessment

This article reports the quality assessment of the LLOD Cloud in terms of accessibility, use of open licenses, and amount of data. The LLOD Cloud[1] counts more than 200 KGs in June 2024. From a preliminary evaluation of linguistic datasets modeled according to the LOD principles and attached to a scientific contribution indexed by Scopus, it resulted that several resources are still missing to the Cloud. In the direction of confirming it, the LLOD Cloud diagram is explicitly declared to be an ongoing project inspired by the LOD cloud diagram authored by Richard Cyganiak and Anja Jentzsch and includes open, available, and interlinked linguistic resources. As a consequence, it is expected that the diagram will incorporate an increasing number of resources over time.

At the current stage, the LLOD Cloud is organized in categories, which are *corpora*, *lexicons&dictionary*, *terminologies, thesauri & Knowledge Base (KB)*, *linguistic data categories*, *linguistic resource metadata*, *typological database (DB)* and *other*. Categories are not balanced, as can be observed in `Column #` in Table 1. Moreover, we can observe the presence of a consistent portion of the Cloud categorized as `Other`. It might raise the question of whether the categorization is adequate and detailed enough. Further studies are required to verify if the current categories are aligned with linguistics' expectations.

Methodology. Data to perform the quality assessment are retrieved by publicly accessible pages attached to resources published within the LLOD Cloud. We have downloaded all the resources in a single JSON file, and we retrieved the `title` of each KG, `keywords` to distinguish the LLOD category used to classify linguistic resources, `sparql` which report the SPARQL endpoint, if any, `full_download` or `other_download` to retrieve any download format attached to each resource, `triples` which model the amount of data, and the `license`. Besides returning the link of the SPARQL endpoint and the download format(s), the LLOD cloud also returns the status of each link. As a result, we can identify all the resources attached to a working link.

Starting from the quality dimensions defined by Zaveri et al. [8], the performed analysis reports the quality assessment of the LLOD Cloud in terms of availability and licensing, belonging to the accessibility dimensions category, and the amount of data belonging to the contextual dimensions category. We consider a resource accessible if it is attached to at least a download format or a working SPARQL endpoint. Scrutinizing all the licenses attached to LLOD, we manually identified all licenses recognized as Open, such as the Apache License[2], the MIT license[3], or the Creative Commons licenses[4]. The amount of data dimension is aligned with the number of triples directly returned from the LLOD Cloud.

Results. The Python script to compute the quality dimensions along with the LLOD.json file are openly and publicly available online[5]. Quality assessment results are reported in Table 1

[1]LLOD Cloud: https://linguistic-lod.org
[2]Apache License: https://www.apache.org/licenses/LICENSE-2.0
[3]MIT License: https://opensource.org/license/MIT
[4]Creative Commons Licenses: https://creativecommons.org/licenses
[5]GitHub repository: https://github.com/isislab-unisa/LLODCloudQuality
Persistent DOI on Zenodo: https://doi.org/10.5281/zenodo.13449868

Table 1
It reports the quality assessment of the LLOD Cloud in terms of accessibility, amount of data and licensing. Results are organized in categories used within the Cloud. Results are sorted according to the number of KGs per each category. Rows are colored only for readability. Legend: **OL** stands for Open License, **A** stands for Accessibility, **SE** stands for a working SPARQL endpoint. The letters reported represent units of data measurement — not numeric quantities.

Category	#	OL	A	SE	OL &A	OL &SE	Data Tot.	A	SE	OL&A	OL&SE
Corpora	73	45	6	0	5	0	771M	744M	0	4M	0
Lexicons & Dictionary	72	34	36	2	8	2	1B	322M	0	4M	0
Other	50	16	25	3	9	0	12B	421M	538M	0	0
Terminologies, Thesauri & KB	15	8	12	4	5	0	31M	28.5M	403K	25M	0
Linguistic Data Categories	14	7	9	2	8	2	788M	83M	8M	67M	8.6M
Linguistic Resource Metadata	2	0	1	0	0	0	988	0	0	0	0
Typological DB	1	0	1	1	0	0	133K	133K	133K	0	0
Overall	227	110	90	12	31	4	15B	13B	430M	894M	8.6M

clustered per LLOD categories. Per each category, we report:

- column `#` - the number of LLOD resources in that category;
- column `OL` - the number of LLOD resources attached to an Open License (OL),
- column `A` - the number of LLOD resources attached to a working download mechanism, including to a working SPARQL endpoint;
- column `SE` - the number of LLOD resources provided with a working SPARQL endpoint, and naturally a subset of the accessible resources;
- column `OL&A` - the number of openly accessible LLOD resources, meaning that they have a working download mechanism and are attached to an OL;
- column `OL&SE` - the number of LLOD resources openly accessible via a working SPARQL endpoint;
- `data` columns reporting the total amount of data (column `Tot.`), the one accessible via a working download mechanism (column `A`), via a working SPARQL endpoint (column `SE`), via a working download mechanism and provided with an open license (column `OL&A`), and via a working SPARQL endpoint and provided with an open license (column `OL&SE`)

Discussion & Conclusive Thoughts. LLOD resources are not uniformly distributed over the categories, as *corpora*, *lexicons&dictionary*, and *other* cover the majority of the LLOD. Less than half of the LLOD resources are attached to an OL, and in some cases, it is even deprecated, as happens by using CC BY-NC 2.0 license[6] instead of the updated 4.0 version. This overall picture

[6]CC BY-NC 2.0: https://creativecommons.org/licenses/by-nc/2.0

is almost coherent with each category, with few exceptions, as the *corpora* category where 60% of resources are attached to an OL, while resources belonging to the *linguistic resource metadata* and *typological DB* categories completely miss OL. 40% of the resources are accessible via a working download mechanism, but only 5% of them are attached to a working SPARQL endpoint. The situation is even worse when we focus on openly accessible LLOD resources, as they drop to 14% while considering any download mechanism, to only four resources if we need an openly accessible LLOD attached to a working SPARQL endpoint. This results in an extraordinary amount of data being left untapped. While naming the worst case, LLOD in the *lexicons&dictionary* category sums up to 1G of triples, while 322MM can be downloaded, only 4MM can be openly and freely reusable, while no data can be accessed via SPARQL endpoints. In summary, there is a huge potentiality in the LLOD regarding the variety of resources and amount of data, but accessibility and the rare use of open licenses hinder their exploitation.

References

[1] C. Chiarcos, S. Hellmann, S. Nordhoff, Linking linguistic resources: Examples from the open linguistics working group, in: C. Chiarcos, S. Nordhoff, S. Hellmann (Eds.), Linked Data in Linguistics: Representing Language Data and Metadata, Springer, Heidelberg, 2012, pp. 201–216.

[2] I. Aldabe, C. Chiarcos, J. Gracia, C. Roeder, M. Villegas, Towards a linguistic linked open data cloud: The open linguistics working group, Linked Open Data–Creating Knowledge Out of Interlinked Data 948 (2012) 19–25.

[3] I. Aldabe, C. Chiarcos, J. Gracia, C. Roeder, M. Villegas, The open linguistics working group, Linked Open Data–Creating Knowledge Out of Interlinked Data 948 (2014) 19–25.

[4] C. Chiarcos, S. Nordhoff, Linghub: A linked data based portal supporting the discovery of language resources, in: Proceedings of the ACL 2012 System Demonstrations, 2012, pp. 81–86.

[5] C. Chiarcos, Linguistic linked open data for speech processing, Journal of the International Phonetic Association 44 (2014) 103–120.

[6] C. Chiarcos, Ll(o)d and nlp perspectives on semantic change for humanities research, Journal of Language Technology and Computational Linguistics 27 (2012) 21–36.

[7] C. Chiarcos, S. Nordhoff, Observing lod: Its knowledge domains and the varying behavior of ontologies across them, Journal of Web Semantics 32 (2015) 18–29.

[8] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, S. Auer, Quality assessment for linked data: A survey, Semantic Web 7 (2016) 63–93. doi:10.3233/SW-150175.