Roadmap and common agenda for future research on linguistic data science

Project Acronym	Nexus Linguarum
Project Title	European network for Web-centred linguistic data science
COST Action	CA18029
Starting Date	26 October 2019
Duration	54 months
Project Website	https://nexuslinguarum.eu
Responsible Author	Jorge Gracia
Contributors	Christian Chiarcos, Milan Dojcinovski, Katerina Gkirtzou, Dagmar Gromann, Fahad Khan, Ilan Kernerman, Penny Labropoulou, Patricia Martín Chozas, Gilles Sérasset, Blerina Spahiu, Dimitar Trajanov, Andon Tchechmedjiev
Version | Status	v1 | final
Date	15 April 2024

Acronyms List

CA	COST Action
CIG	COST Innovators Grant
LD	Linked Data
LLD	Linguistic Linked Data
LLM	Large Language Model
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
LR	language resource
MC	Management Committee
MoU	Memorandum of Understanding
NLP	Natural Language Processing
W3C	World Wide Web Consortium
WG	Working Group

Table of Contents

EXECUTIVE SUMMARY

This report on a “roadmap and common agenda for future research on linguistic data science” has been the result of a long reflection process by the NexusLinguarum community along the project duration. The first part of this paper is dedicated to the identification and discussion of a series of challenges in the field of Linguistic Data Science (LDS), more particularly in Linguistic Linked Open Data (LLOD). They comprise: entry barriers to the technology, sustainability, coverage of current representation models, metadata, cross-lingual linking, under-resourced languages, and multilinguality. Then, a possible roadmap is proposed to address such challenges and progress towards an ideal ecosystem for LLOD. A dedicated discussion on the relation between LLOD and the emergent Large Language Models (LLMs) is also provided. Finally, a concrete plan to continue the activities of the NexusLinguarum COST Action, with the idea of continuing progressing along the proposed roadmap, is provided.

1. Introduction

This report has been the result of a long reflection process by the NexusLinguarum community along the project duration. It summarises the outcomes of some dedicated meetings devoted to identify challenges and define a roadmap and future steps for the community such as:

- "CLARIN Café on Linguistic Linked Data" (29 April 2021, online)
- "Roadmap" session, at the NexusLinguarum 5th plenary meeting (7-8 September 2023, Milan)
- "Day of W3C language technology community groups" at LDK'23 conference (12 September 2023, Vienna)
- "Roadmap with a common agenda for future research on linguistic data science. Sustainability plan" at the NexusLinguarum 6th plenary meeting (21 March 2024, Athens)

Furthermore, a substantial part of this report largely relies on a published journal article on the more general topic "Multilinguality and LLOD: A Survey Across Linguistic Description Levels" (Gromann et al. 2024), which contains dedicated sections to analyse challenges and an envisioned ecosystem for the Linguistic Linked Open Data (LLOD) field, and that has been authored by members of NexusLinguarum. Finally, some other elements have been taken from the NexusLinguarum deliverable on "Guidelines and Best Practices on Linguistic Linked Open Data" (Martin-Chozas et al., 2024).

2. Challenges

As explained in the Memorandum of Understanding (MoU[1]) of the NexusLinguarum[2] COST Action (CA 18209), LLOD constitutes a key technology in the area of Linguistic Data Science. In this section we analyse the main challenges that this technology is facing. First, we show the results of an informal but illustrative survey on how LLOD is perceived by members of the community; secondly, we describe its main challenges based on both the personal experience of experts in the field and on the analysis of the scientific literature.

2.1. Questionnaire

In order to have a more grounded view of how LLOD is perceived by the members of the Action and what their main expectations and identified challenges are, we conducted an informal survey during a "roadmap" session at the NexusLinguarum 5th plenary meeting (survey conducted on 7/09/2023, with 38 respondents). The results are the following:

Degree of awareness of the LLOD cloud raised by NexusLinguarum

Question 1: "Before Nexus, were you aware of the LLOD cloud?"

[1] https://e-services.cost.eu/files/domain_files/CA/Action_CA18209/mou/CA18209-e.pdf

[2] https://nexuslinguarum.eu/the-action/

Answers:

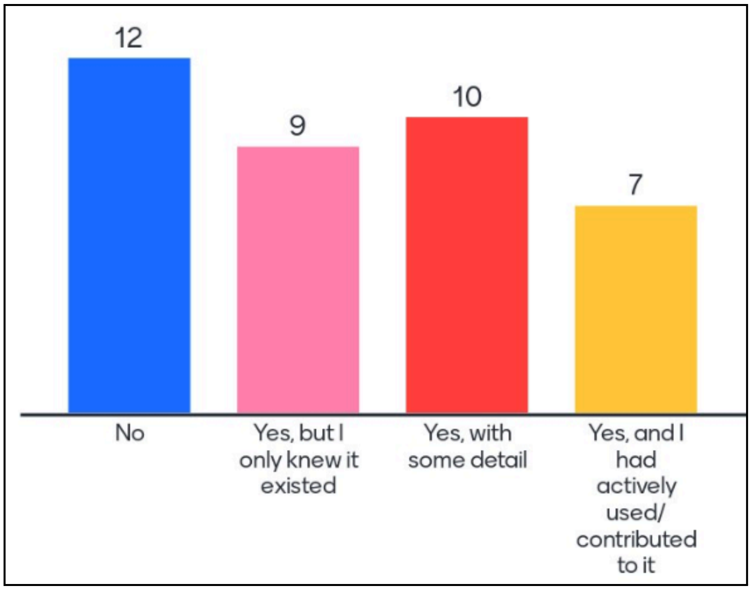

Figure 1: Answers to Question 1 (“Before Nexus, were you aware of the LLOD cloud?”)

Question 2: “After Nexus, your degree of awareness of the LLOD cloud is…”

Answers:

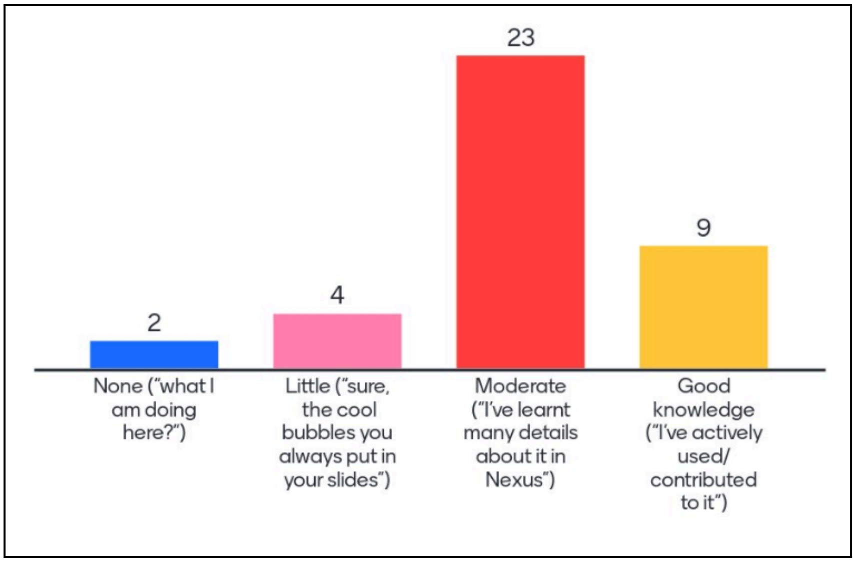

Figure 2: Answers to Question 2 (“After Nexus, your degree of awareness of the LLOD cloud is…”)

Issues and preferences with regard to the LLOD cloud

Question 3: “If you were a data consumer of the LLOD cloud, you would prefer…”

Answers:

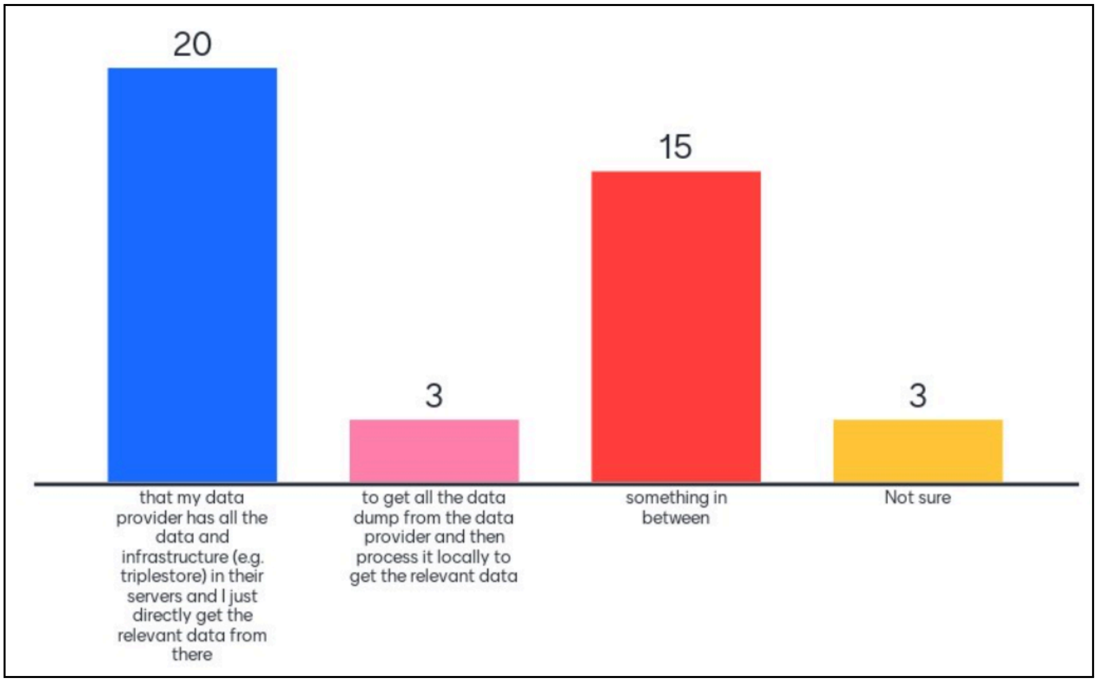

Figure 3: Answers to Question 3 (”If you were a data consumer of the LLOD cloud, you would prefer…”)

Question 4: “If you were a data provider of the LLOD cloud…”

Answers:

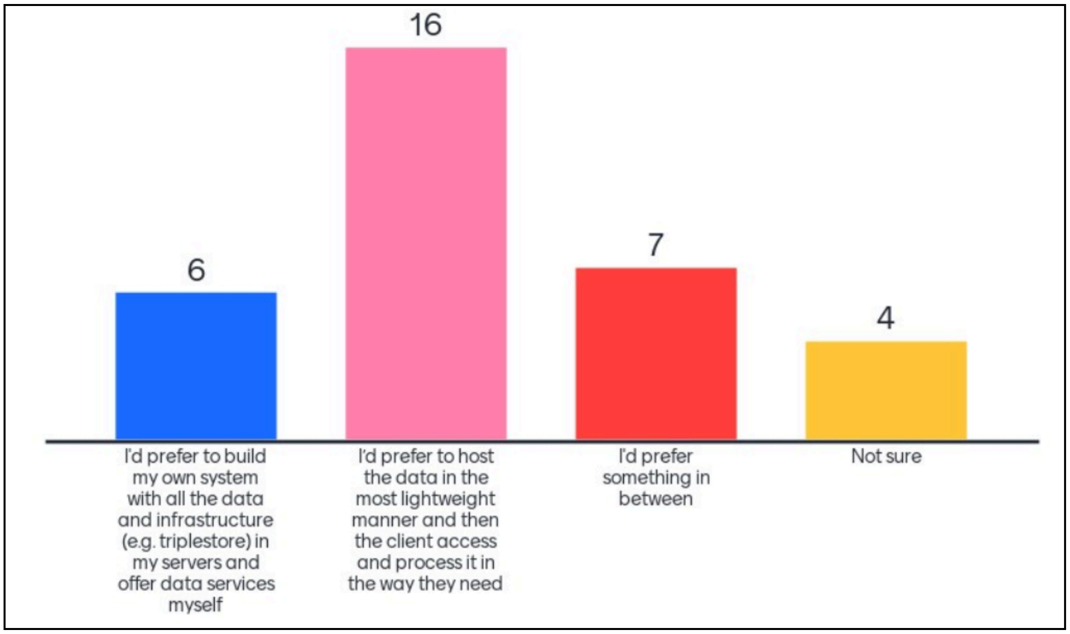

Figure 4: Answers to Question 4 (“If you were a data provider of the LLOD cloud…”)

Question 5: “RANK the following issues of LLOD according to importance”

Answers:

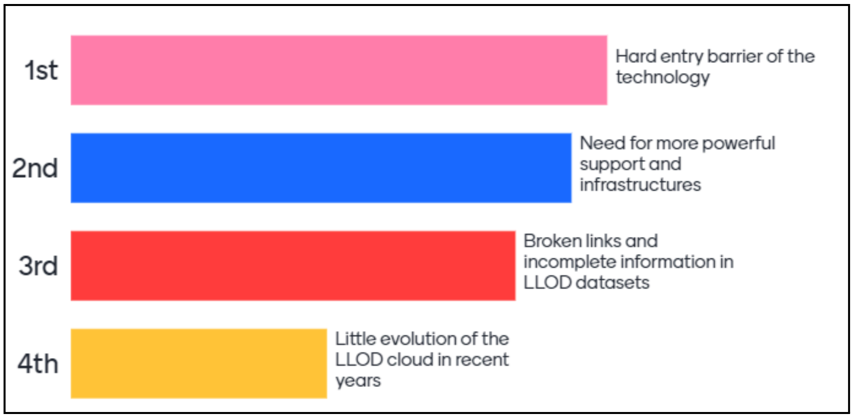

Figure 4: Answers to Question 5 (“RANK the following issues of LLOD according to importance”)

The results show how the majority of respondents moved from none or very little knowledge of LLOD before NexusLinguarum into a great majority who is using it with at least moderate familiarity (Questions 1 and 2).

As for preferences (Questions 3 and 4), potential data consumers prefer that the provider has the data and infrastructure in their servers so they just get the relevant data from there, while potential data providers would prefer to host the data in a lightweight manner, delegating the accessing and processing of the data to the data consumer. This result seems to lead to two confronting solutions. Thus, to accommodate everybody's preferences some hybrid or intermediate approaches will be needed.

Finally, among some known issues of the technology, the respondents identified the “hard entry barrier of the technology” as the most prominent one, closely followed by the “need for more powerful support and infrastructures” (Question 5).

2.2. Identified challenges

The rest of this section summarises the analysis conducted by Gromann et al. (2024) as part of the activities of Working Group 3 “Support for linguistic data science” of NexusLinguarum.

Despite its rising popularity and recognition of its usefulness by different disciplines, the LLOD Infrastructure has some new and old challenges to overcome (cf. Gracia et al. 2012, Chiarcos et al. 2020, Declerck et al. 2020).

As a result of our systematic study, and also based on our own experience, we analyse in this section a number of such challenges to be addressed in order to bring LLOD to its full potential for representing and linking multilingual language data across linguistic levels. Although some of these challenges are common to LD in general (e.g. sustainability), we do not want to miss the opportunity to refer to them here because they are also crucial for the LLOD community. Other issues related to language resources or linguistic data in general, but not so much specific to LD or LLOD (e.g. legal issues, ownership, data protection), are out of the scope of this section.

Entry Barriers to the Technology

One of the central challenges revolves around enabling researchers and practitioners, who may not be familiar with the LLOD framework, to utilise it effectively. As with any emerging technology, LD presents a steep learning curve, requiring proficiency in RDF, OWL, SPARQL, and specific models such as OntoLex-Lemon. Furthermore, new adopters will need certain technical support to set up the appropriate infrastructure, which may vary depending on their needs, from simple storage of RDF dumps to fully-fledged triple stores with de-referenceable mechanisms.

Another challenge results from the amount of language resources that are available, which increases the complexity of issues related to interoperability. In fact, once a resource in the LLOD cloud is discovered, its access and exploitation are not always straightforward. Additionally, the presence of abandoned resources and broken links in the LLOD cloud might be a discouraging experience for newcomers.

To address these challenges, it is not only imperative to develop tools and standards and to conduct research, but also to invest in education by means of training schools and courses. These educational activities are critical for the continued growth and advancement of the LLOD infrastructure and the expanding LLOD community. In that respect, ongoing research projects and networks, and the activities of several W3C community groups, are progressing in that direction. For instance, NexusLinguarum has organised a series of training schools around the topic of linguistic linked data, and has supported a number of tutorials and seminars on this topic. Additionally, Linghub, developed in the context of the LIDER[3] and Prêt-à-LLOD[4] projects, aims at alleviating the issue of discoverability and reusability of language resources, by indexing a large amount of language resources metadata in a way that can be easily exploited by software agents as well as by humans.

However, there is still a need for user-friendly visual interfaces and working environments for working with LLOD (frameworks such as VocBench (Stellato et al., 2020) are a step in the right direction), as well as tools and infrastructures for an easier deployment of (linguistic) semantic data on the Web.

Researchers and practitioners who specialise in specific linguistic description levels and actively generate linguistic resources covering one or more linguistic description levels are not necessarily LLOD-savvy. Lowering the LLOD entry barrier is in the interest of the LLOD community as well as of such researchers and practitioners. For the former, it is important to increase the coverage especially of yet under-represented linguistic description levels, such as phonetics and phonology, pragmatics, dialogue, sign languages, and diatopic representations. For the latter, it is of interest to maximise re-usability and interoperability of their often manually curated resources. Finally, addressing these challenges will contribute to lowering the entry barriers for both the LLOD community and the researchers and practitioners specialising in specific linguistic description levels.

[3] http://lider-project.eu/

[4] https://pret-a-llod.eu/

Sustainability

Ensuring the sustainable hosting of RDF data exposed as linked data on the web is another critical challenge, not limited to LLOD but common to LOD in general. This challenge involves balancing the efforts between data providers, data consumers, data hosts, language resource providers, technology developers, and linked data application developers. As it has been recently reported in several fora[5] and scientific papers (Chiarcos 2021), there is a need for sustainable hosting solutions for the RDF data exposed as linked data on the Web. The main issues, which are common not only to LLOD but to LOD in general, are:

1. Data consumers may want content negotiation mechanisms and server side infrastructure (triple store + SPARQL endpoints). This can be a burden on the host/provider.
2. Alternatively, the burden can be put on data consumers, if they need to download and locally process RDF data dumps.

Focusing on the federation and queryability of linked data resources, a scenario that is ideal from the perspective of the user would be if the host can expose the data via a SPARQL endpoint -- which could be directly queried by a client without setting up a local infrastructure.

On the other hand, real-world infrastructures currently allow only to deposit data as files with the media types plain/text (plain text) or application/octet-stream (arbitrary binary data). To use this data as RDF, an application needs to guess the correct format and, in many cases, it has to download all data first and set up a local query engine.

One compromise between both extremes is to deposit data as uncompressed files with appropriate RDF-compliant media types (e.g., text/turtle, application/ld+json, etc.), with a small additional burden on the data provider and host to indicate the proper media type, e.g., by means of content negotiation (Chiarcos 2021). Then, the data can just be imported into an RDF triple store (or a SPARQL web service) by means of the SPARQL keywords LOAD or FROM.

On a technical level, some other intermediate solutions have been proposed, such as:

- Linked data Fragments[6] is an effort to redistribute the load between clients and servers by means of the Triple Pattern Fragments.
- SPARQLer[7] is a web service that allows running queries against external data sets that can be consulted using the SPARQL FROM keyword. SPARQLer is just a blank installation of Apache Jena[8] with permissions granted to eliminate the need for a user to set up a local RDF database.
- RDF-HDT[9] is a community standard for binary compressed RDF data that can be directly queried by means of SPARQL. HDT requires downloading external data, but does not require setting up a local SPARQL endpoint.

More powerful support and infrastructures are, however, still needed. Something analogous to WordPress[10] for websites, but for small linked data providers.

[5] https://www.clarin.eu/event/2021/clarin-cafe-linguistic-linked-data

[6] https://linkeddatafragments.org/

[7] http://www.sparql.org/

[8] https://jena.apache.org/

[9] https://www.rdfhdt.org/

[10] https://wordpress.com/

Some steps in this direction are Databus[11], TriplyDB[12], and Semantic media wiki[13]. We consider that larger infrastructures, like the European Language Grid (ELG[14]), CLARIN[15], or the European Language Data Space[16], can play an active and important role here.

Coverage of current representation models

To lower the entry barrier to the LLOD cloud, a representation mechanism for linguistic data is crucial. While most linguistic description levels are well-represented in the current landscape, some areas, such as phonetics and phonology, pragmatics, dialogue, sign languages, and diatopic representations, lack comprehensive LLOD models. These gaps present challenges not only for the LLOD community but also for researchers and practitioners specialising in these areas. For the latter group, maximising the reusability and interoperability of their manually curated linguistic resources is essential.

One level that encompasses more facets in linguistic research than LLOD representations currently provide is phonetics and phonology. PHOIBLE 2.0 provides a very large cross-linguistic inventory of phonemes in more than 2,000 languages[17]. However, it is one of the few LLOD models for this description level available and many areas from socio-phonetics to phonetics in language acquisition might require a dedicated representation. Areas such as sign phonetics from a multilingual perspective, not solely focusing on a specific sign language, and representing sign languages as LLOD resources, in general, are yet to be explored systematically. Regarding the level of pragmatics, there are some models, such as the OLiA discourse extension, that focus on representing dialogue structure, however, this linguistic research field has more to offer, e.g., speaker attitude, turn taking, etc.

Another important aspect of representing linguistic data as linked data is the ease to move across and between distinct description levels. Fortunately, interoperability is one of the key assets of the LLOD concept. One predominant approach of the LLOD community that becomes evident in this survey is the extension of existing representation models with dedicated modules for specific levels. For instance, numerous extensions to OntoLex-Lemon and OLiA provide a communal base representation to which to link specific information, e.g., phonetic features and morpho-syntactic annotations across languages. Models with different theoretical underpinnings can equally and jointly be explored by means of their linked representation in the LLOD cloud. However, this brings us back to the ease of access to LLOD resources, which is a requirement to be attractive to a wide audience. Only then it is feasible to explore cross-disciplinary linguistic research in multiple natural languages.

When it comes to specific language resources, especially corpora, formalisms such as POWLA have been proposed a decade ago, but still very few primary corpus data or corpus metadata have been published in the LLOD cloud. This raises the question of whether there is a need to extol the virtues of querying, consistency controlling, and linking such data, also to other types of resources and across languages, more explicitly, or whether the entry barriers to the LLOD cloud and/or representation models is too high for providers of such data. Within NexusLinguarum

[11] https://databus.dbpedia.org/

[12] https://triply.cc/

[13] https://www.semantic-mediawiki.org/

[14] https://www.european-language-grid.eu/

[15] https://www.clarin.eu/

[16] https://language-data-space.ec.europa.eu/

[17] https://phoible.org/

there has been an initiative to collect feedback from corpus providers on the use of LLOD in this context. Despite the results not being conclusive yet, they indicate that large national corpus providers tend to be reluctant to utilise linked data, if they had even heard about it, stating that resources tend to be unstable (without automatic redirects if a resource fails), that it is hard to integrate linked data with current machine learning methods, and that there is a lack of tutorials for LLOD Infrastructures.

These arguments suggest that the reluctance to publish corpora as linked data is more an issue of LOD Infrastructure, which needs to become more stable, easy-to-use, and ideally integrated with state-of-the-art machine learning methods, than with proposed representation models. Nevertheless, this survey article shows that some representation models have been taken up more vibrantly than others, which might not necessarily allow conclusions about the model itself but rather constitutes a call to the LLOD community to interact and collaborate more closely with communities that curate multilingual data. For instance, strong showcases of performing multilingual linguistic research on an easily accessible LLOD Infrastructure might help the case.

To conclude, lowering the entry barrier to LLOD is in the interest of both the LLOD community and these domain-specific researchers and practitioners. Expanding coverage, especially for under-represented linguistic description levels, is vital.

Metadata

Metadata provides a challenge for a broad audience involved in linguistic research, language resource creation and curation, phonology, translation, and related fields, all of whom can benefit from improved metadata standards and linked data solutions. One remarkable issue when publishing LRs on the Web is that their metadata is scattered across the different language repositories, which makes it problematic to ensure effective search procedures across the repositories. Furthermore, there are different standards adopted for different repositories, which makes data accessibility and linking problematic. There are also difficulties in harmonising metadata from different repositories in order to provide a single point of access to search for relevant language resources across repositories.

Actually, linked data provides suitable mechanisms to solve such issues. In this regard, we advocate for an increased use of agreed vocabularies for LRs metadata description, such as the Meta-Share OWL ontology (McCrae et al. 2015). Other types of metadata that might be of interest for the LLOD cloud is the Information Coding Classification (ICC), or the licensing information in machine-understandable ways. To overcome existing inconsistencies among different language resources, di Buono et al. (2022) propose a promising methodology for fixing and enriching metadata for LOD Cloud and Annohub repositories.

Besides metadata for the description of language resources, metadata for the development of particular use cases in linguistics also poses interesting challenges. In fact, means to represent information on discourse structures and discourse relations in a multilingual setting and pragmatics in general is currently poorly represented in LLOD, as are phonetics and phonology. One especially challenging aspect within the context of LLOD is that all these metadata need to be linked to the participant in a specific study rather than to a language resource or a data repository. Thereby, LLOD could support the development of meta-analysis studies, e.g., to analyse the development of a specific grammatical element across studies. Furthermore, as studies on translation inference in general and in relation to pragmatics have shown, the potential to query data inventories in a structured manner with a specific research question in mind across languages, potentially even from a diachronic perspective, opens up entirely new research avenues for different linguistic branches. For phonology, for instance, such interlinking holds the potential to analyse speech patterns across a large number of languages and representation modes.

Cross-Lingual Linking

Cross-lingual linking enhances the efficiency and effectiveness of multilingual data integration and knowledge sharing. Thus, it is beneficial for NLP and Semantic Web researchers, cross-cultural studies, ontology development, benchmark creation, language resource provision, and language technology development, among others.

Interlinking multilingual resources is not straightforward since when entities are described in different natural languages, string similarity measures cannot be applied directly. This task poses several challenges: (1) the structure of graphs can be different and the structure-based techniques will not be of much help; and (2) even if the structures are similar to one another, the properties themselves and their values are expressed in different natural languages.

From the perspective of conceptualisation, other issues arise in the linking task (Gracia 2012a): (a) conceptualisation mismatches due to language and cultural discrepancies; (b) conceptualisation mismatches due to the perspectives from which the same domain is approached; or even, (c) different levels of granularity in the conceptualisation. Despite the recent advancements in the field, all the referred issues remain valid and give room for further research.

Another remarkable challenge is the need of benchmarks to support the evaluation of methods and algorithms on cross-lingual linking, in a Semantic Web context. Current efforts in that direction are the Multifarm track, which is part of the periodic Ontology Alignment Evaluation Initiative (OAEI[18]), and the Translation Inference Across Dictionaries (TIAD[19]) shared task. The Multifarm dataset is composed of the alignments among seven ontologies of the Conference domain, translated into eight different languages, thus resulting in 45 different language pairs that serve as a gold standard for cross-lingual ontology matching systems. Despite its obvious interest, this dataset only covers one specific domain. More domains and languages would be necessary to further stimulate the progress in the field. Additionally, the TIAD task has been beneficial and led to progress in the field of cross-lingual linking. However, this is specific to a concrete task, which is bilingual lexicon induction, and measures performance among three language pairs (French, English, Portuguese) only. A broader language coverage and the extension of this idea to similar tasks involving cross-lingual link discovery would be also beneficial.

Under-Resourced Languages

The main challenges that under-resourced languages face can be grouped into two[20]: technological barriers (e.g., lack of the large amounts of data needed to support current deep learning approaches) and cultural and socio-economic barriers (e.g., the low number of language resources hinders cultural heritage maintenance). There are a good number of ongoing efforts and initiatives aimed at the promotion of languages that are often under-resourced. However, the resulting data remain in project-specific formats, leading to insufficient data access, possibilities

[18] http://oaei.ontologymatching.org/

[19] See latest campaign description at https://tiad2022.unizar.es/

[20] See https://nexuslinguarum.eu/wp-content/uploads/2022/10/02_Policy-Briefs.pdf

for sharing and integration for query and comparison. In that context, linked data arises as a natural solution to address this scenario, providing mechanisms for interoperability at a Web scale.

There are some remaining open issues in the application of LD to under-resourced languages, though, like the necessity of modelling languages that are very rich morphologically and the still low adoption of LLOD at the morphological level. A second remarkable issue, as pointed out by Gillis-Webber and Tittel (2019), is the current limitation of language tags when dealing with very specific language variants or dialects. The latter is, however, not an LLOD-specific issue, but something broader that involves internationalisation of the Web at a larger scale. Nevertheless, potential solutions to that issue might come in LD-native ways following the example of lines of works such as Lexvo.org[21], a database that brings information about languages, words, characters, and other human language-related entities in a linked data format.

Another category of under-resourced languages that is important to consider is that of Sign Languages. As Sign Languages require multimodal representation, they provide a particularly interesting challenge for representation models. Since Sign Languages are not organised the same way as spoken languages, representing them might require additional elements of current formats for spoken and written languages. In fact, while current resources cover movements of hands and body in images for a sign, information on mouthing or mouth movements are missing among other types of information. Even if this information was available for many signs, there are only a few fully annotated corpora of a decent size. Within European projects, such as Intelligent Automatic Sign Language Translation (EASIER[22]), Sign Language Translation Mobile Application and Open Communications Framework (SignON[23]), and the COST Action NexusLinguarum, work has been done to improve this. For instance, Declerck et al. (2023) utilise the Open Multilingual Wordnet (OMW) infrastructure[24] as a pivot between sign language data, i.e., in German, Greek, English, and Dutch[25] with extensions to Danish, Icelandic, and Swedish Sign Sanguages, and propose OntoLex-lemon as a format for interlinking and aligning sign and spoken language resources. A hurdle while doing so is that the concepts expressed in Sign Languages and Spoken Languages may differ largely. For several iconic signs, for example, a distinguishing expression in the surrounding Spoken Language may not exist.

Multilinguality

Multilinguality plays a crucial role in enhancing access to linguistic data across various languages, making it a valuable source for linguists, entities dedicated to language preservation and revitalization, multilingual communication organisations, language resource curators, and Semantic Web researchers. The Semantic Web in general, and linked data in particular, has been repeatedly identified as a core technology to overcome language barriers on the Web (Gracia et al. 2012), since it has mechanisms to represent, traverse, and integrate, data in different languages, mediated by a common ontological layer. However, the main question is whether LLOD has really helped in making the Semantic Web more multilingual. Studies indicate that the

[21] http://lexvo.org/

[22] https://www.project-easier.eu/

[23] https://signon-project.eu/

[24] https://omwn.org/

[25] Both Dutch as used in the Netherlands (NGT) and Dutch as used in Belgium (VGT) The spoken language is largely the same, the signed languages are really different languages.

number of language tags used in the Semantic Web increased, but the dominance of English never stopped (di Buono et al. 2022).

In terms of comparison of the LLOD cloud and the broader LOD one, one wonders if LLOD is more “multilingual” than the general LOD. The current availability of linguistic data in the LLOD cloud in terms of languages needs a more systematic exploration.

There is also a need to focus on the coverage and details on the granularity of available data (lexical entries / links to other languages through translation of common referents / availability of data from the different linguistic description levels / etc.). An “observatory” would be needed to measure the quality and evolution of linguistic data along such dimensions.

3. Roadmap

In a previous analysis, one decade ago, Gracia et al. (2012) studied the challenges posed by the so-called Multilingual Web of Data and proposed a roadmap towards its full realisation. In a first stage, they proposed the development of new (lightweight) representation models along with simple techniques for ontology localisation, cross-lingual querying and linking. The idea was to ensure early adoption of LLOD and provide the required incentives for the development of more complex infrastructures in future stages. In a second stage, semantic search engines might index multilingual lexical information available on the Web and support answering ad hoc queries in any language. More complex models and services would be developed in this second stage, supporting cross-lingual natural language processing applications requiring deeper multilingual lexical knowledge. Finally, the third stage would be more user-centred, with people more motivated to provide multilingual lexical information. An ecosystem of services would be available for cross-language querying, on-demand translation, cross-lingual mappings, etc. Search engines might be able to process natural language questions in any language and adapt their result presentation to conventions of the linguistic and cultural community to which the user belongs.

As our literature analysis attests, there has been substantial progress in the field over the last ten years. However, this progress did not always move in the direction predicted in the mentioned roadmap. Some goals have been accomplished, to judge from the emergence of new models (e.g., lexicog[26]) and updated versions of other well-established ones (e.g., Lemon[27]), as well as the (still moderate) progress in cross-lingual link inference (e.g., the TIAD campaign). However, the roadmap envisioned a more central role for the final Web user, who’d be more aware of the incentives and rewards that publishing linguistic information as LD should bring. We are still far from that. Recent progress has been achieved mainly in academic contexts, for specialised studies with specialised linguistic data. This is not bad in itself, of course, and there are very successful stories in the application of LLOD for linguistic research (e.g., the LiLa[28] project). However, some pieces are still missing for a larger uptake of the LLOD technologies. For instance, a major role of semantic search engines, as envisioned in the 2012 roadmap, or a higher level of infrastructural/sustainability support, as reported in Section 2.

[26] https://www.w3.org/2019/09/lexicog/

[27] https://www.w3.org/2016/05/ontolex/

[28] https://lila-erc.eu/

3.1. Towards an Ideal Ecosystem for LLOD

In the rest of this section, we propose a new roadmap with the next steps that the community might take to address the challenges reported in Section 2, in order to attain an ecosystem of truly interoperable linguistic data on the Web, multilingual in nature, across different linguistic levels. These steps are not intended to be sequential and can overlap.

1. Step I. More robust and sustainable open infrastructures should be in place, to support small and medium scale data providers who cannot afford their own hosting infrastructure. Since the technology is already in place, this is a matter of promoting its adoption and carrying out new national and international LD projects with a clear focus on infrastructure development. In parallel, more educational efforts are needed to make the advantages of LLOD visible to a new generation of researchers and practitioners. While this step is a general LOD issue, it is of crucial importance to achieve a highly Multilingual LLOD cloud as this necessarily requires publishing many datasets of varying size and language coverage from many publishers who cannot afford their on-premise infrastructure.
2. Step II. New models, along with new systems for RDF generation and linking, will be developed to cover linguistic description levels currently under-represented in the LLOD cloud. This will enable truly cross-disciplinary linguistic research in multiple natural languages, at Web scale.
3. Step III. Development of an "observatory" to measure the quality and evolution of linguistic data on the Web along several dimensions (language, linguistic level, usage, etc.). Stable metadata models and repositories will be in place, with the ultimate aim of not only discovering relevant language resources, but really accessing their data and enabling their direct re-use and inter-operation. Metadata models are of tremendous importance in Semantic Web and LOD in general. Their usage is, however, mainly disregarded in the NLP community. This step is the key towards usages where the required resources would be automatically discovered and used in the LLOD, rather than fixed (and usually imported) at development time.
4. Step IV. Massive population of the LLOD cloud with the maximum possible number of languages (thousands better than hundreds) and resources. That will create a critical mass of data to be eventually exploited by final language applications. This should cut the vicious circle resulting in lack of data caused by lack of exploitation opportunities and vice-versa.
5. Step V. Development of a fully fledged family of services for easy upload and integration of multilingual linguistic data on the Web, language independent access and querying of linguistic data, and seamless integration of such data with NLP services and tools. That will also include user interfaces for browsing/editing linked data.

3.2. Linguistic Linked Data and Language Models

Finally, as an extra step that might be traversal to the above five, an extensive study on the inter-relation between linguistic linked data and the emergent Transformer-based architectures is required, to better place our technologies in the current scientific and technical landscape. We foresee such interaction in two directions (see Figure 5): (i) how Language Models can enrich LLOD, by applying techniques such as relation discovery, translation, ontology population, etc., and (ii) how LLOD can enrich Language Models, enhancing domain adaptation of language models and attaining better and more explainable results.

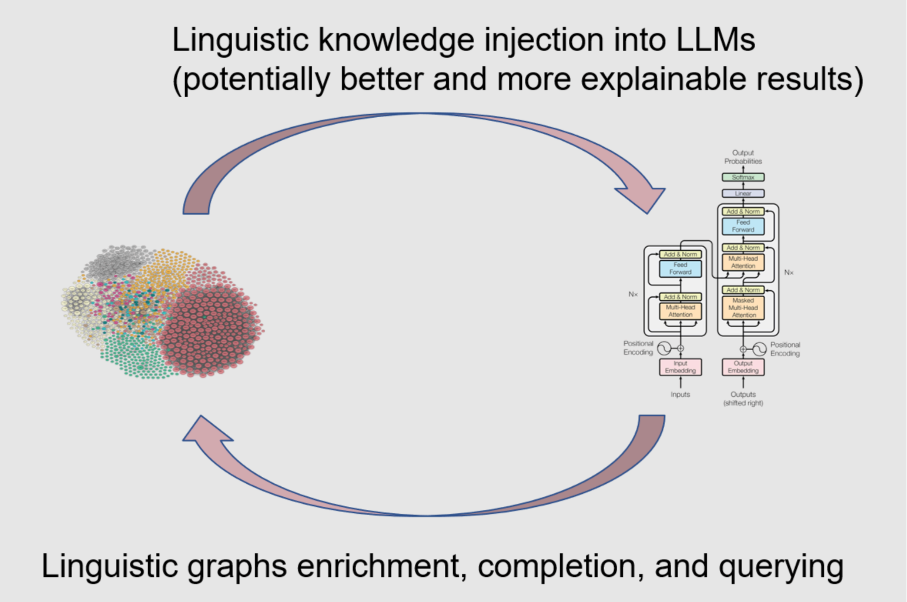

Figure 5: relation between linked data and contemporary language models

Some practical realisations of this research line are already happening nowadays. For instance to generate SPARQL queries from natural language with LLMs, to generate RDF from natural language sentences, or to easily convert between formats, e.g. RDF/XML to TTL. Some potential applications in both directions are:

From LLMs to LLD

- **Semantic Annotation**. LLMs can help in annotating text data semantically, saving time and reducing the manual effort.
- **Support dictionary/lexicon creation**. LLMs can be used to generate domain-specific dictionaries, lexicons and/or ontologies in specific fields.
- **Automated Mapping.** LLMs can help in mapping unstructured or semi-structured text to structured ontologies, thereby aiding the seamless integration of disparate data sources.

- **Query Handling**. LLMs can interpret natural language queries and translate them into formal queries that can be executed on structured linguistic databases, enhancing user accessibility.

From LLD to LLMs

- **Fact-Checking**. LLD can serve as an external knowledge base for fact-checking, enabling LLMs to validate the information they generate and provide more accurate results, while decreasing the computing effort and cost.
- **Citation Support**. With access to LLD, LLMs can provide more authoritative responses by citing reliable sources.
- **Bias Mitigation.** By incorporating balanced and verified information, LLD can assist in mitigating the inherent biases that might exist in LLMs.
- **Transparency and Explainability**. LLD can facilitate the generation of explanations for the LLM's output, thereby improving transparency and trust.

4. Sustainability plan

As NexusLinguarum comes to an end, and to make the aforementioned roadmap a reality, we need to ensure that all that we have built within Nexus perdures and takes a life of its own, continuing in a sustainable manner even without the tools provided by COST. The cornerstone of our sustainability strategy lies in establishing or reviving community groups and efforts that would pursue the efforts durably in time. The NexusLinguarum sustainability plan comprises the continuation of the Action's activities in various fronts:

- **COST Innovators Grant**. We submitted a proposal for a COST Innovators Grant (CIG), with a focus on increasingly involving industrial actors and creating the foundation for long-term financing. The proposal was titled "NexSus: Sustaining NexusLinguarum" and is currently under evaluation by the COST Association. The CIG is aimed to build bridges between research and take-up at market, product, service, or societal level in the field of Linguistic Data Science. CIGs have a duration of 12 months and offer the possibility to create additional impact during the year after the end of the Action and they benefit from the same networking activities available to COST Actions.
  Even if the CIG is not funded, many elements proposed therein could be supported through other community-led initiatives and by future project proposals that fund the development of specific aspects, much like the funding schemes used in many European infrastructures.

- **W3C groups.** As discussed by Martin-Chozas et al. (2024), the BPMLOD W3C community group will sustain the creation of guidelines and their regular updates and evolutions, while the Ontolex-Lemon W3C community group will drive the development and adoption of community-wide standards by using BPMLOD recommendations as a springboard to bootstrap the development of said standards. While these communities can certainly continue functioning without any kind of financial support, we will strive to create a business plan that also ensures financial sustainability.

- **Erasmus Mundus Joint Master.** One of the main goals of the Action was to work out a curriculum for a Europe-wide master degree that the participating institutions could adopt to train a new generation of researchers in the area, thus introducing Linguistic Data Science in a cross-discipline academic infrastructure (see Costa et al, 2024). As result of the Action's activities in this regard, a proposal was elaborated for a European Master in Linguistic Data Science (EMLDS), a 120 ECTS English-language programme awarding a Joint Degree. Based on the collaboration amongst four European universities (NOVA University of Lisbon, Università Cattolica del Sacro Cuore of Milan, University of Zaragoza, and University of Galway) and their associated partners, the European Master in Linguistic Data Science aims at providing a novel approach to linking Linguistics to Computer Sciences and Data Science in terms of its methodology, scientific content, pedagogical approach, and curricular structure. The EMLDS proposal, submitted on 14/02/2024, is currently under evaluation by the European Commission.

- **LDK conference.** The Language Data and Knowledge (LDK) international conference series is a biennial conference series on matters of human language technology, data science, and knowledge representation, initiated in 2017 and supported by an international Scientific Committee of leading researchers in natural language processing, linked data and Semantic Web, language resources and digital humanities. Following the success of its first two editions in Galway, Ireland (2017) and Leipzig, Germany (2019), the Action, in common agreement with the LDK Scientific Committee, decided to support and fund the continuation of this conference series. Its third edition took place in Zaragoza (2021) and the fourth one in Vienna, Austria (2023), both of them organised under the umbrella of NexusLinguarum and COST. In fact, the goals of the LDK conference are very well aligned with the Action topics. LDK has served as the "flagship" conference for NexusLinguarum. Currently, the 2025 edition of LDK is in preparation. Both the chair (Jorge Gracia) and vice-chair (Dagmar Gromann) of the Action are acting as principal chairs of the conference. This new edition falls outside the duration of the project, but will continue acting as an invaluable meeting point for the community and for the exchange of ideas.

- **Other workshops**. A good number of workshops were supported by NexusLinguarum in different venues. A substantial number of them were born within the Action, as a result of the joint effort and initiative of working group members. It is expected that they will remain active beyond the duration of the Action, and will serve as yet another forum where its objectives and results can be discussed and projected over time. A non-exhaustive list of such workshops born in NexusLinguarum is:
    - Sentiment Analysis & Linguistic Linked Data (SAALD)
    - Taxonomy and annotation of offensive language

    - Discourse studies and linguistic data science (DisLiDas)
    - PROfiling LINGuistic KNOWledgE gRaphs (ProLingKNOWER)
    - Linking Lexicographic and Language Learning Resources (4LR)
    - TermTrends: Models and Best Practices for Terminology Representation in the Semantic Web
    - DLnLD: Deep Learning and Linguistic Linked Data

- **NexusLinguarum as an informal online community**. During the final Management Committee (MC) meeting of the Action (21/03/2024 in Athens, Greece), it was decided that, in addition to the aforementioned continuation mechanisms, NexusLinguarum will continue part of its activities as an informal online community. Such a community will be composed initially by the Action participants, during at least 5 years after the end of the Action or until a proposal to acquire another status (e.g., a non-profit organisation) coming from any former member(s) of the Action will get the consensus of the NexusLinguarum community. During this time, Universidad Politécnica de Madrid (UPM, current Action Grant Holder institution) will continue hosting the Action website and mailing lists, and will take the temporary ownership of the NexusLinguarum logo and brand. The MC delegated into the Action's core group the task of elaborating some minimal statutes for the NexusLinguarum online community.
  The idea is to continue using the NexusLinguarum website for online presence, the current mailing lists for communication, the github account for the development of code and data, and the slack channel for quick communication.

5. Conclusions

This document presents a "roadmap and common agenda for future research on linguistic data science". A number of challenges on this topic were identified during the NexusLinguarum COST Action, some of them already addressed by the Action's activities, while others remained open. Addressing them is the basis of a roadmap plan also documented in this report. Finally, a sustainability plan was proposed to continue the activities of NexusLinguarum beyond the end of the project.

We hope that the analyses and reflections contained in this document serve as a guide for future research in the field of LDS and can serve to motivate future research projects and proposals related to the topics and objectives of NexusLaguarum.

Bibliography

Chiarcos, C., Klimek, B., Fäth, C., Declerck, T., & McCrae, J. P. (2020), "On the Linguistic Linked Open Data Infrastructure", in: Proceedings of the 1st International Workshop on Language Technology Platforms, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 8–15. ISBN 979-10-95546-64-1. https://www.aclweb.org/anthology/2020.iwltp-1.2

Chiarcos, C. (2021). "Get! Mimetypes! Right!", 3rd Conference on Language, Data and Knowledge (LDK 2021) (Vol. 93, p. 5:1-5:4). Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing. https://doi.org/10.4230/OASICS.LDK.2021.5

Costa, R., Gracia, J., Amaro, R., Carvalho, S., McCrae, J., Passarotti, M., Recharte, S. (2024) "Academic Common Curriculum on Linguistic Data Science - LDS", NexusLinguarum D3.3 Deliverable, https://nexuslinguarum.eu/wp-content/uploads/2024/04/Deliverable-D3.3-common-curriculum-1.pdf

Declerck, T., McCrae, J., Hartung, M., Gracia, J., Chiarcos, C., Montiel, E., Cimiano, P., Revenko, A., Sauri, R., Lee, D., Racioppa, S., Nasir, J., Orlikowski, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M. F., Khvalchik, M., Gonzalez, M., & Cooney, K. (2020). "Recent Developments for the Linguistic Linked Open Data Infrastructure". *Proc. of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 5660–5667. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.695.pdf

Declerck, T., Bigeard, S., Khan, F., Murtagh, I., Olsen, S., Rosner, M., Schuurman, I., Tchechmedjiev, A., and Way, A., (2023) "A Linked Data Approach for linking and aligning Sign Language and Spoken Language Data", in: Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages, European Association for Machine Translation, Tampere, Finland, 2023, pp. 11– 21. https://aclanthology.org/2023.at4ssl-1.2

di Buono, M. P., Gonçalo Oliveira, H., Barbu Mititelu, V., Spahiu, B., & Nolano, G. (2022). "Paving the way for enriched metadata of linguistic linked data". *Semantic Web*, *13*(6), 1133–1157. https://doi.org/10.3233/SW-222994

Gillis-Webber, F. and Tittel, S. "The Shortcomings of Language Tags for Linked Data When Modeling Lesser-Known Languages" (2019), in: 2nd Conference on Language, Data and Knowledge (LDK 2019), (OASIcs), Vol. 70, 2019, pp. 4:1–4:15. ISSN 2190-6807. ISBN 978-3-95977-105-4. doi:10.4230/OASIcs.LDK.2019.4.

Gracia, J., Montiel-Ponsoda, E., & Gómez-Pérez, A. (2012). Cross-lingual linking on the multilingual web of data (position statement). *Proc. of the 3rd Workshop on the Multilingual*

Semantic Web (MSW 2012) at ISWC 2012, Boston, USA, 936. http://ceur-ws.org/Vol-936/paper6.pdf

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Gómez-Pérez, A., Buitelaar, P., & McCrae, J. (2012). Challenges for the multilingual Web of Data. Journal of Web Semantics, 11, 63–71. https://doi.org/10.1016/j.websem.2011.09.001

Gromann, D., Apostol, E.-S., Chiarcos, C., Cremaschi, M., Gracia, J., Gkirtzou, K., Liebeskind, C., Mockiene, L., Rosner, M., Schuurman, I., Sérasset, G., Silvano, P., Spahiu, B., Utka, A., Truica, C.-O., & Oleškevičienė, G. V. (2024). "Multilinguality and LLOD: A Survey Across Linguistic Description Levels". Semantic Web Journal.

McCrae, J. P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., & Cimiano, P. (2015). "One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web". *Proc. of 12th Extended Semantic Web Conference (ESWC 2015) Satellite Events, Portorož, Slovenia, 9341*, 271–282. https://doi.org/10.1007/978-3-319-25639-9_42

Martín-Chozas, P., Dojchinovski, M., Gkirtzou, K., Khan, A.F., & Tchechmedjiev, A. (2024) "Guidelines and Best Practices on Linguistic Linked Open Data" NexusLinguarum D1.4 deliverable. https://nexuslinguarum.eu/wp-content/uploads/2024/03/Deliverable-D1.4-Guidelines-and-Best-Practices-on-LLOD.pdf

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., & Keizer, J. (2020). VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons. *Semantic Web*, *11*(5), 855–881. https://doi.org/10.3233/SW-200370