Proceedings of the Linked Data in Linguistics Workshop @ LREC2022, pages 69–77
Marseille, 24 June 2022

A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data

Anas Fahad Khan[1], Christian Chiarcos[2], Thierry Declerck[3], Maria Pia di Buono[4], Milan Dojchinovski[5,6], Jorge Gracia[7], Giedre Valunaite Oleskeviciene[8], Daniela Gifu[9,10]

[1] Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Italy, fahad.khan@ilc.cnr.it
[2]Applied Computational Linguistics, Goethe University, Frankfurt, Germany, chiarcos@cs.uni-frankfurt.de
[3]DFKI GmbH, Saarland Informatics Campus, Saarbrücken, Germany, declerck@dfki.de
[4]UNIOR NLP Research Group, University of Naples "L'Orientale", Italy, mpdibuono@unior.it
[5]Faculty of Information Technology, Czech Technical University in Prague, milan.dojcinovski@fit.cvut.cz
[6]DBpedia Association/InfAI, Leipzig University, Germany, dojchinovski@informatik.uni-leipzig.de
[7]Aragon Institute of Engineering Research (I3A), University of Zaragoza, Spain, jgracia@unizar.es
[8]Institute of Humanities, Mykolas Romeris University, Vilnius, Lithuania, gvalunaite@mruni.eu
[9]Faculty of Computer Science, Alexandru Ioan Cuza University of Iasi, Romania, daniela.gifu@info.uaic.ro
[10]Institute of Computer Science, Romanian Academy - Iasi Branch, Romania, daniela.gifu@iit.academiaromana-is.ro

Abstract

This article discusses a survey carried out within the NexusLinguarum COST Action which aimed to give an overview of existing guidelines (GLs) and best practices (BPs) in linguistic linked data. In particular it focused on four core tasks in the production/publication of linked data: generation, interlinking, publication, and validation. We discuss the importance of GLs and BPs for LLD before describing the survey and its results in full. Finally we offer a number of directions for future work in order to address the findings of the survey.

Keywords: Linguistic Linked Data, Guidelines, Best Practices

1. Introduction

This article has its origin in a survey on the use of specific Linked Data (LD) vocabularies for different categories of language resources. The survey was carried out by the COST Action "CA18209 - European network for Web-centred linguistic data science"[1]. At the time of writing this article, the Linguistic Linked Open Data (LLOD) Cloud,[2] consisting of datasets belonging to the linguistics domain, makes up one of the largest subsets of the Linked Open Data (LOD) cloud, with 227 datasets out of a total 1301 in the whole LOD cloud[3]. However, after over a decade of LLD, and despite the advantages and opportunities of LLD, there is still room for improvement, not least in terms of languages covered[4] and types of linguistic dataset represented in the LLOD cloud. The provision of clearly formulated guidelines and best practices written in different languages (and featuring use cases dealing with a range of languages) and types of resources could help to close these gaps and make Linguistic Linked Data (LLD) more accessible. In addition, such a documentation can also help in the exploitation of the datasets in the LLOD cloud, i.e., to help it realise its full potential. We therefore reflect in this paper on the current state of such guidelines and best practises, the main topics they cover, their targeted audience, etc. as well as their limitations and the aspects that future materials of this type should fulfil. Indeed, there exist caesurae and weaknesses in the available documentation which prevent the full exploitation of LD principles for linguistic data. The remainder of the paper is organised as follows. In Section 2 we identify some desirable aspects that guidelines and best practises on LLD should fulfil. Then, Section 3 describes our survey of currently available materials and, finally, Section 4 features a discus-

[1]The short name of the COST Action being "NexusLinguarum", https://nexuslinguarum.eu/

[2]The LLOD cloud is accessible at http://linguistic-lod.org/llod-cloud

[3]https://lod-cloud.net

[4]According to a recent policy brief (Bosque-Gil et al., 2021) on under-resourced languages the availability of language resources demonstrates tremendous differences across languages. Some languages like English have an abundance of the resources available for LLOD technologies, while some other languages show scarcity of resources. This lack of language resources is damaging for at least two reasons: first, the application of advanced data processing technologies is limited as they require extensive data and next, the automated development and enrichment of language resources becomes really scarce. In addition, linguistic resources vary in their depth of the information available on numerous linguistic features, thus the use and re-use of the data poses a twofold challenge of processing in width and in depth of the available linguistic resources. The resources are unevenly developed in different languages and some languages may not have the material developed available for LLOD technologies.

sion of what is missing, with a number of suggestions addressed to the LLD community to produce useful guidelines and best practices.

2. The Role of Guidelines and Best Practices in LLD

2.1. Some Definitions

One aim of the survey was to give an overview of existing guidelines (GLs) and best practices (BPs) with respect to four core tasks in the production/publication of linked data. These are: **generation**, **interlinking**, **publication**, and **validation**. The survey also helped in determining what is missing or needs to be updated in those areas, leading to the intention to work on these gaps, also in collaboration with other initiatives. Before we continue, however, we should clarify what we mean here by 'guidelines" and "best practices" In the first instance, we can adopt the definition given by the Cambridge English Dictionary[5], stating that a *guideline* is:

> *information intended to advise people on how something should be done or what something should be.*

For *best practice*, we can adopt the Merriam-Webster definition[6]:

> *a procedure that has been shown by research and experience to produce optimal results and that is established or proposed as a standard suitable for widespread adoption.*

Understood in this way, there are relatively few resources which can label themselves either as guidelines or best practices, or anything that could be construed as a synonym of these, in the context of Linguistic Linked Data (LLD). But there is a reasonably large number of other types of material and resources which fulfil, in part, the role of a set of guidelines and best practices as we have defined them above. These include, for instance, one or more sections in the technical report for a standard or individual chapters in an introductory textbook. Our survey therefore took into consideration all of these types of material and resources. We describe our methodology, data gathering process and results in Section 3.

2.2. Desiderata

Understanding the advantages of LLD and the many opportunities it offers as a means of publishing linguistic data as FAIR data[7] requires some level of technical appreciation of the Semantic Web, of RDF and other formalisms as well as a number of other technologies. Nonetheless, in order to increase the uptake of LLD amongst non-specialists, it is essential that materials are made available which are accessible to non-specialists and which give clear instructions and ways of doing common tasks (the role of GLs/BPs). A related issue here is the need for LLD specific technologies which target non-specialists (as opposed to more generic Semantic Web oriented applications and technologies such as protégé[8])[9]. The use of more accessible tools will in turn make the production of more accessible guidelines more viable; something we discuss in Section 4.

In other cases, the provision of clear and easy-to-understand guidelines have been essential in helping to introduce standards and technologies to target audiences. This, for instance, is the case with the Text Encoding Initiative Guidelines[10], which in addition to describing the Text Encoding Initiative approach to annotation themselves (and the elements of which it consists) also incorporates a valuable introduction to XML itself targeted towards humanists. In this context, therefore, there is no clear line between what counts as didactic materials and guidelines and best practises; this is why we have included two self-contained online courses in our list of miscellaneous materials in Section 3.3.

As in any other domain, the use of GLs/BPs in LLD helps to fill the gap between a technical description of a standard and its use in practice; and indeed both kinds of documentation help to ensure the interoperability, and therefore FAIRness of resources[11]. However, it takes on a special significance for LLD given that Linked Data is one of the core technologies which is helping to make FAIR a reality. We end this section with a list of desiderata for LLD GLs/BPs based on the experience of the authors as both consumers and compilers of such documents:

- Multilinguality: they should not just be in English, but should make LLD accessible to speakers of other languages;
- They should be easy to find and access, preferably with an open licence and not behind a paywall; this very fits in the spirit of LL**O**D;
- They should give clear instructions for how to carry out different tasks and be as self-contained as possible (and save users from having to wade through text that is not relevant for their information need). In particular, they should be organised according to the task they are developed for;
- They should be pitched at different levels of expertise but especially for beginners (given we need to

[5]https://dictionary.cambridge.org/dictionary/english/guideline
[6]https://www.merriam-webster.com/dictionary/best%20practice
[7]https://www.go-fair.org/
[8]https://protege.stanford.edu/
[9]There are few generally accessible tools that offer specific provision for LLD use cases, one of those that does exist is VocBench, see (Stellato et al., 2020)
[10]https://tei-c.org/Guidelines/
[11]https://www.go-fair.org/fair-principles/

increase uptake of the technology);

- They should cover (at least) the types of resources listed in the LLOD cloud, and the four tasks of generation, interlinking, publication, and validation;
- They should be aware of existing tools which can be integrated in the workflow
- Be regularly updated to ensure they keep up to date with the latest technology/models/tools.

This list of desiderata will help us evaluate the already existing materials which we have found in our survey and which we look at in the following section, as well as to suggest what to prioritise when it comes to producing new materials.

3. A Survey of Already Available Materials

In order to come up with a candidate list of resources for our survey, we solicited input from the members of the NexusLinguarum COST Action, a group that consists of researchers and linguistic linked data experts with extensive experience in numerous relevant projects and initiatives. In addition, this work also benefited from the extensive process of data collection which was carried out as part of the survey paper on LLD models (Khan et al., 2022 in press) produced as part of Task 1.1 of the NexusLinguarum Action; this included the compilation of a survey of LLD-relevant projects and other relevant initiatives (i.e. W3C community groups). Each of the resources contained in the survey have been described/categorised using a number of salient metadata fields. The fields were chosen with an eye to the potential (re-)usability of these resources. Accordingly, we have specified the level of expertise which is assumed by each resource according to the following categorisation. Note that the "Beginner level" of expertise assumes some basic knowledge of linked data and the Semantic Web, e.g., the concept of a triple, the fact that linked data is structured as a series of subject-object-triples what a SPARQL endpoint is. We do not deal with basic materials for learning about linked data and the Semantic Web here, since our focus is on linguistic linked data and not linked data in general. However, the beginner level of expertise should not assume any specialist knowledge of different areas of (Computational) Linguistics or NLP. For instance, materials which required an intermediate level of familiarity of corpus linguistics but only a basic level of familiarity with (linguistic) linked data would be classed as "Intermediate". An "Intermediate level" of expertise in this context assumes either an intermediate level of familiarity with LLD and/or with some area of Computational Linguistics. The "Advanced level" of expertise is defined similarly.

Target Audience	Description
Beginner	Assumes little or no LLD or technical knowledge
Intermediate	Assumes some LLD or technical knowledge
Expert	Assumes advanced LLD or technical knowledge

Table 1: Levels of Expertise.

Additionally, in our survey we have listed a number of keywords for each resource, including the tasks it is useful for and the kind of resource it covers. In the latter case, we have taken the classification used to categorise the resources in the LLOD cloud, namely (abbreviations in parentheses are used in the survey tables below): Corpora **(Corp)**; Lexicons and Dictionaries **(LD)**; Terminologies, Thesauri and Knowledge Bases **(TTKB)**; Linguistic Resource Metadata **(LRM)**; Linguistic Data Categories **(LDC)**; and Typological Databases **(TD)**. In addition, whenever a resource assists in carrying out one or more of the four tasks which we are focusing on in this deliverable, i.e., generation **(Gen)**, interlinking **(InL)**, publication **(Pub)** or validation **(Val)**, we also add it as a keyword. Note that **Gen** here also includes the sub-tasks of data modelling and conversion of datasets into LLD. In the following subsections, we look at the different kinds of materials described in the survey[12].

3.1. Guidelines and Best Practices

In this section, we consider GL/BP's that either advertise themselves as such or that very clearly have this purpose, that is, the provision of guidelines and best practices for LLD, as a primary aim (as distinct e.g., from technical reports for standards or textbooks which, while fulfilling the role played by GLs and BPs, also have other, distinct aims). It became clear during the information gathering phase of this survey that there was a dearth of materials or resources fitting this description. Here we can, however, mention two different sets of materials, the first of which was produced as a result of work carried out by the now dormant 'Best Practices for Multilingual Linked Open Data' (BPMLOD) W3C community group, and the second of which was an output of the LIDER project [13].

Table 2 describes the eight guidelines made available as part of the BPLMOD set of guidelines. These comprise guidelines for generating multilingual[14] and bilingual[15]

[12]The full survey will be made available as a NexusLinguarum deliverable in April 2023.

[13]https://lider-project.eu

[14]http://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/

[15]https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/

dictionaries, wordnets[16], TBX terminologies[17], developing NIF services[18] and LLOD aware servicess[19] and creating corpora with NIF[20]. Finally, there are guidelines for LLD exploitation[21]. It is notable that all the BPMLOD guidelines are from 2015, seven years from the time of writing and prior to the new version of *lemon*, OntoLex-Lemon, which was published in 2016[23]. This is problematic because there are numerous classes and properties which exist in OntoLex-Lemon and not in *lemon* and vice versa. It is also prior to the publication of the OntoLex-Lemon lexicographic module in 2019[24] (something which clearly affects the first two dictionary related guidelines). As well as being out of date, they do not cover all tasks and all types of resources. This is problematic, given the lack of alternative and more recent materials.

Table 3 summarises the eight reference cards which were made available by the LIDER project. These include guides to publishing linked data[25], language resource licensing[26], inclusion in the LLOD cloud[27], data IDs[28], language resource discovery with Linghub[29], NIF corpora[30], the representation of crosslingual links[31] and language resource documentation in datahub[32].

Such cards are structured as 'how to' instructions to address different types of target audiences, e.g., data publisher, data creator, and different scopes, e.g., publishing LD on the Web. Furthermore, they clearly state the steps and the knowledge needed, e.g., RDF knowledge, together with the resources/tools useful for reaching the goal.

These reference cards were intended to offer sets of guidelines for carrying out a number of tasks, ranging from publication, adding metadata, and including resources on the LLOD cloud, which were accessible for beginners. Again all of these cards date from a specific year, and once again this year is 2015 (an exemplary year for LLD guidelines and best practices!). Unfortunately, we were unable to find any licensing information for these reference cards, so it is unclear how and when they can be re-used. Note also that neither the BPMLOD guidelines nor the reference cards deal directly with the validation of linked data, nor do they offer any special assistance in the case of working with typological databases. Additionally, the reference cards run to two pages each and are limited in the amount of information they offer with respect to the task of enriching a linguistic dataset with metadata or dataset crosslinking.

Finally, the *lemon* cookbook[33], which was an output of the Monnet project[34] which provided an introduction to the *lemon* model, describing each of its submodules and generally fulfilling the role of a set of guidelines. For Ontolex-Lemon, the official W3C community report of the final specifications of the model fulfils the role played by the *lemon* cookbook for Ontolex-Lemon as we discuss in Section 3.2.

3.2. Standards

Another group of documents relevant to this discussion are technical reports and specifications for LLD-related standards. These include 'official' formal standards: those that are issued and maintained by designated institutions[35] and subject to a formal, institution-specific process of proposal, review, revision, confirmation and withdrawal. These can be subject to a number of constraints on formats and means of presentation that usually make them less accessible than some other kinds of materials we've looked at above and which take a more didactic stance. In addition to formal standards, a number of specifications exist, which are treated as *de facto* standards specifications by the community without being published as official standards by some standardisation body. In what follows we largely focus

[16] http://bpmlod.github.io/report/WordNets/index.html (Unofficial Draft)

[17] https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/

[18] https://www.w3.org/2015/09/bpmlod-reports/nif-based-nlp-webservices/

[19] http://bpmlod.github.io/report/LLOD-aware-services/index.html

[20] http://bpmlod.github.io/report/nif-corpus/index.html (Unofficial Draft)

[21] https://www.w3.org/2015/09/bpmlod-reports/lld-exploitation/

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

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

[25] http://bpmlod.github.io/report/LLOD-aware-services/index.html

[26] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/How-to-publish-linguistic-linked-data-Reference-Card.pdf

[27] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Inclusion-in-the-LLOD-Cloud-Reference-Card.pdf

[28] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/DataID-Reference-Card.pdf

[29] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Discovering-Language-Resources-with-Linghub.pdf

[30] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/NIF-Corpus-reference-card.pdf

[31] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/How-to-represent-crosslingual-links-Reference-Card.pdf

[32] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Documenting-a-language-resource-in-Datahub.pdf

[33] https://lemon-model.net/lemon-cookbook/index.html

[34] https://cordis.europa.eu/project/id/248458

[35] These include standardisation bodies such as, for example, W3C, OASIS and ISO.

Title	License	Target	Keywords	Last Updated
Guidelines for Linguistic Linked Data Generation: Multilingual Dictionaries (BabelNet)	W3C Community FSA[22]	Expert	babelnet, lemon, wordnet, generation, **LD**, **Gen**	2015
Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries	W3C Community FSA	Expert	bilingual dictionary, lemon, translation, multilingual lexical resources, **LD**, **Gen**	2015
Guidelines for Linguistic Linked Data Generation: Multilingual Terminologies (TBX)	W3C Community FSA	Expert	Multilingual terminologies, TBX, resource conversion, **TTK**, **Gen**	2015
Guidelines for Developing NIF-based NLP Services	W3C Community FSA	Expert	NIF, NLP services	2015
Guidelines for LLD exploitation	W3C Community FSA	Intermediate	LLD services, use cases	2015
Guidelines for Linguistic Linked Data Generation: WordNets	W3C Community FSA	Expert	wordnet, lemon, **LD**, **TTK**	2015
Guidelines for Linked Data corpus creation using NIF	W3C Community FSA	Expert	NIF, **Corp**	2015
Guidelines for LLOD aware services	W3C Community FSA	Expert	LLD services, use cases	2015

Table 2: The BPMLOD Guidelines

on such community standards since there do not exist many LLD-specific (as opposed to linked data specific) formal standards. In fact, we look at individual technical reports and specifications for such *de facto* standards to see the extent to which such documentation can fulfil the role of GLs and BPs[36]. The primary purpose of such documentation is undoubtedly to give an exhaustive and unambiguous description of a standard. In many cases, however, they are also intended to assist users in applying the standard, often by providing examples of its use in typical use cases – and in this they play the same role as GLs/BPs.

Standards for lexical-semantic resources We will start by looking at the specifications for OntoLex-Lemon, a W3C community standard for lexical resources which we mentioned above and which was originally inspired by the UML-based proprietary ISO standard Lexical Markup Framework (LMF) on its first iteration. The OntoLex-Lemon specifications were published in 2016 in the W3C namespace as a community report by the W3C Ontology-Lexicon group[37]; note however that as reported in document itself OntoLex-Lemon is not a W3C recommendation (and neither is it on the W3C recommendations track). Besides detailed descriptions of the single classes and properties in the model, the specifications also give (simple and fairly accessible) examples of the use of the latter, both in the form of diagrams and snippets of code. In general the text of the specifications is fairly expansive and goes beyond the more technical presentation of, e.g., ISO standards making the document accessible to the Beginner level of user. These specifications can therefore be said to fulfil the role of a set of beginner's guidelines or a primer to OntoLex-Lemon. On the other hand, there are many use cases (especially for generation but also other tasks) which they don't (and given their status as guidelines shouldn't) capture. Moreover, the guidelines are so far only available in English with the examples mostly in English, in addition to a handful of others in Latin, French, Spanish and German. As well as the necessity of translations of the OntoLex-Lemon specifications in other languages

[36] We leave out documentation for vocabularies like SKOS, SKOS-XL, DCAT, DCMI which aren't LLD specific even though they are often used in LLD datasets, neither do we deal with LLD (de facto) standards which do not currently have accessible reports/specifications.

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

Title	License	Target	Keywords	Last Updated
How to publish Linguistic Linked Data	N/A	Beginner	Linking Data, Resolvable URIs, **Gen**, **Pub**, **InL**	2015
Language Resource Licensing - ODRL Reference Card	N/A	Beginner	RDF Conversion, Data Modeling, Linking Data, Resolvable URIs, **LRM**, **Gen**, **Pub**	2015
Inclusion in the LLOD Cloud	N/A	Beginner	LLOD Cloud, Datahub, Linked Dataset, **Pub**	2015
Data ID	N/A	Beginner	Dataset description, DataID, Resource Metadata, Pub	2015
Discovering Language Resources with Linghub	N/A	Beginner	LingHub, Resource Discovery, Language Resources	2015
NIF corpus	N/A	Beginner	NIF, RDF, Corpus Conversion, **Corp**	2015
How to represent crosslingual links	N/A	Beginner	Cross-lingual Linked Data links, Cross-lingual mapping, **Pub**, **InL**	2015
Documenting a language resource in Datahub	N/A	Beginner	Metadata documentation, DataHub, DCAT, data description, **LRM**, **Gen**, **Pub**	2015

Table 3: Lider Project Reference Cards

(with examples in other languages too), it is also clear that we need more Intermediate and Expert level materials dealing with more advanced modelling topics for OntoLex-Lemon. However, it is not unreasonable to assume that the popularity of OntoLex-Lemon is due in no small part to the accessibility of the specifications, both in terms of the fact that they are openly available and unlike ISO standards like LMF aren't closed or behind a paywall, and their readability.

Two years after the publication of these specifications, the W3C Ontology Lexicon group published the specifications for an extension to the OntoLex-Lemon model, dealing this time with lexicographic resources, namely, the OntoLex-Lemon Lexicography Module (lexicog) [38]. In line with the specifications of the original model, these specifications were furnished with illustrative examples for individual classes and properties. The limitations of these guidelines are the same as those of the original model; as will likely be the case for another two follow-up OntoLex-Lemon modules in an advanced phase of preparation (the first dealing with the representation of morphology, the second with frequency, attestation and corpus data), with others also being planned, including an extension for terminologies (this would make a good start in developing guidelines for the TTK category).

Standards for linguistic annotation There is currently no settled consensus as to which is the most suitable linguistic annotation mechanism for LLD. This is important since linguistically annotated data plays a vital role in current NLP/AI technologies.[39] NLP Interchange Format NIF and the Web Annotation standard, a W3C recommendation that developed out of the Open Annotation community. NIF is a community standard developed in a series of research projects at the AKSW Leipzig, Germany, and still maintained by that group. In addition to that, it enjoys a semi-official status as a component of the Internationalization TagSet (ITS 2.0) which is a formal W3C standard that describes the application of NIF. Web Annotation is a W3C recommendation that evolved out of the Open Annotation vocabulary, a community standard originally published

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

[39] Here the importance of collaborating with small and medium enterprises (SMEs) in the development of new standards should be emphasised. This would have the effect of helping them to establish new business relationships and enter new markets early. Vice versa, the experiences of SMEs in working with Semantic Web technologies would likely prove crucial to strategic discussions about the Web's future.

as a community report of the W3C Community Group Open Annotation.

Both Web annotation and NIF build on the use of URIs (resp., IRIs) for addressing corpora, and this coincides with the use of URIs (IRIs) in TEI and XML standoff formats. A typical UR/IRI consists of two main components, a base name that serves to locate the document, and an optional fragment identifier. For numerous media types and different file formats, different fragment identifiers have been defined, often as best practices (BPs; also referred to as Requests for Comments, RFCs) of the Internet Engineering Task Force (IETF). Other, format-specific standards include the W3C standards SVG (Scalable Vector Graphics),[40] XPointer (for addressing XML documents)[41], or Media Fragments[42]. None of these are specific to linguistic annotation, but they can be used in conjunction with Web Annotation or NIF. The level of presentation in these standards and community standards is relatively technical, its content is normative and oriented towards engineers that are responsible for implementing the corresponding reference functions. None of these standards is particularly user-friendly. In addition to standards and community standards, a URI schema for Web Annotation selectors is provided as a working note that accompanies the W3C recommendation. Again, this document has the same level of technicality. It is therefore clear that this is one area where there is a real necessity for documentation that provides clear GL's and BP's.

3.3. Miscellaneous

Finally, we round off this current section by looking at other types of materials or resources which have served, or which might serve, to play the role of GLs and BPs for LLD, alongside a range of other didactic or expository tasks. One category of materials which can often play this role is textbooks and monographs and here in particular we can cite the introductory text *Linguistic Linked Data: Representation, Generation and Applications*, (Cimiano et al., 2020). This book is intended to be primarily introductory, but also contains intermediate and advanced materials. Although designed to be self-contained, it recommends, in each chapter, a number of additional readings to complete the given overview and to get deeper into some details. The book is structured in four main blocks: preliminaries (a basic introduction to linked data and linguistic linked data), modelling (lexical data, annotated texts, linguistic annotations, metadata), generation and exploitation (generation of LLD resources, linking, workflows), and use cases (multilingual wordnets, digital humanities, discovery of language resources). Although not conceived as a set of guidelines in itself, it shares many commonalities with our previous definition of guidelines, and is a valuable source of reference for those interested in LLD in general or in any of its particular aspects. Overall, there are at least a couple of major drawbacks to using such books as sources for GLs/BPs. For a start, and given current publishing practices (and notwithstanding a growing movement towards publication of open edition) their digital editions tend to be paywalled, with the kind of copyright licenses that mean that the information in them can't be shared – at least not legally. More generally, information contained in them and which pertains to GLs/BPs tends not to be in a self-contained format. There are many similar issues with articles (paywalls, copyright, less focus on providing self-contained sets of GLs/BPs). Another category of material or resource that is salient to the current discussion are didactic or course materials. In order to respond to the information needs of users looking for GLs/BPs these should be self-contained (and not depend on other materials) as well as, preferably, made freely available. Although one can often find slides (both from courses and from conference/workshop presentations) which will in many cases answer specific questions, it's difficult to find materials which can more generally take on the function of GLs/BPs. Here, however, we can mention two courses published on the DARIAH-CAMPUS platform (the latter being as the name suggests an initiative of the DARIAH infrastructure) and which were produced as an output of the ELEXIS European Project and which fulfil in large part the role of GLs/BPs. The first is the course *Modeling Dictionaries in OntoLex-Lemon*[43]; the second is the *Lexicography in the Age of Open Data* [44]. These are much closer to the materials we looked at in Section 3.1, especially the BP-like content of the Lider reference cards.

3.4. Observations

Returning to the list of desiderata listed in Section 2 and in light of the last few sections, what observations can we make with respect to what exists? The most obvious one is simply that there aren't enough materials available fulfilling the role of GLs/BPs for linguistic linked data, and moreover a lot of what exists hasn't been updated for years and doesn't reflect the latest developments in the field. And this is true of all levels of expertise. In the case of OntoLex-Lemon and its extension(s), these are well served by their specifications; moreover, OntoLex-Lemon is regarded as *the* de facto standard for lexicons and dictionaries. This makes it much easier to produce further materials, at least in contrast to cases where there is no such settled standard (or when there are too many incompatible standards). This would argue in favour of initiatives for consolidating competing standards or rendering them interopera-

[40] https://www.w3.org/TR/SVG11/linking.html

[41] https://www.w3.org/TR/xptr-framework/

[42] https://www.w3.org/TR/media-frags/

[43] https://elexis.humanistika.org/resource/posts/modeling-dictionaries-in-ontolex-lemon

[44] https://elexis.humanistika.org/resource/posts/lexicography-in-the-age-of-open-data

ble might[45]. In the case of books and articles, these can be helpful in providing sets of GLs and BPs, but such materials are usually not published as open source publications, or digital editions are behind paywalls, and might not be the organised in a way that's convenient for those searching for specific GLs/BPs. All of which suggests a real need for new GLs and BPs.

Finally, the question of the languages in which GLs/BPs are written in (as well as the kind of examples which they feature) is a crucial one, especially for the uptake of LLD standards and technologies. The lack of information available in languages other than English reflects a similar disparity in language resources. As suggested in the introduction, the provision of GLs/BPs in other languages and/or with the inclusion of a wider range of linguistic examples from typologically diverse languages could help to improve this situation. Overall, the need to provide easy-to-read guidelines and goal-oriented instructions, addressing different levels of expertise and use cases, calls for a re-organisation and integration of existing documentation.

4. Conclusion: What is to be done?

After laying out the current situation with respect to GLs and BPs for LLD, we suggest a number of future work directions. We propose to promote and/or (wherever possible) implement these work directions within the framework of the Nexus Linguarum COST action in collaboration with other initiatives and projects as discussed below.

Update existing GLs and BPs; Solicit feedback for new GLs/BPs Perhaps the lowest hanging fruit here: Given the continuing existence of the W3C BPMLOD group (even if currently inactive), one obvious proposal would be for Nexus Linguarum participants to work with that group on updating already existing GLs. In addition, suggestions for new GLs and BPs could be solicited both from that group and other relevant W3C groups such as the W3C Ontology Lexicon group and Nexus Linguarum mailing lists, and indeed any other relevant community list. This brings us onto our next proposal.

Use case/example driven GLs and BPs; Bridging GLs and BPs and tools As we have seen, there is a real need to adapt and extend GLs and especially BPs with more use case driven examples. One idea would be to reinstate something like lemon patterns, or to make use of a repository of ontology design patterns (this idea is further discussed in (Khan et al., 2022 in press)). In addition, where possible, GLs and BPs should focus on actual implementation of the particular task using a concrete tool or software.

A Central Hub for GLs and BPs. Another proposal would be to establish a central hub for LLD. This would significantly help with the discovery of relevant materials. Currently, there is a lack of a reference point for search and discovery of BPs and GLs.

Open, editable and collaborative GLs and BPs. In order to keep materials up-to-date, it is necessary to enable users to directly contribute to the materials and provide updates when necessary. This can be achieved by providing the materials through a wiki system or using markdown documents. Both this and the previous proposal could be undertaken in collaboration with infrastructures like CLARIN or DARIAH (as part of the Social Sciences and Humanities Open Cloud(SSHOC) cluster[46]).[47]). It should not be neglected that some replication even with stable and well maintained infrastructures might be considered. In fact, one of the past initiatives of DARIAH was to enhance communications between five European Research Infrastructures (ERICs) in the Social Sciences & Humanities (SSH): CLARIN, DARIAH, European Social Survey (ESS), Survey of Health, Ageing and Retirement in Europe (SHARE), Consortium of European Social Science Data Archives (CESSDA). In addition, Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies (PARTHENOS) supports the work of CLARIN and DARIAH. **Interactive BPs and GLs.** While most available materials are static (e.g. PDF documents or static HTML pages) making use of video clips and quizzes would significantly help with knowledge transfer and increase user engagement. In particular, organising a massive open online course (MOOC) could help to deliver learning content online in an interactive way. Different levels could be offered catering to users with different levels of expertise and/or different backgrounds In fact, OER (Open Education Resources) would be more appropriate to cover a wide range of online learning formats, like the ones already mentioned and many more.

5. Acknowledgements

This article is based upon work from COST Action NexusLinguarum – "European network for Web-centered linguistic data science" (CA18209), supported by COST (European Cooperation in Science and Technology) `www.cost.eu`. The article is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015), by the I+D+i project PID2020-113903RB-I00, funded by MCIN/AEI/10.13039/501100011033, by DGA/FEDER, and by the *Agencia Estatal de Inves-*

[45] Indeed, an initiative is underway for such a consolidation for RDF vocabularies for linguistic annotation within Nexus Linguarum

[46] https://sshopencloud.eu/

[47] Indeed in addition to the infrastructures mentioned above there are several other European initiatives supporting the development of a unique platform to access language technologies and tools for all European languages, e.g., European Language Grid https://www.european-language-grid.eu/ (ELG) which could also collaborate in the development of shared and user-oriented documentation.

tigación of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

6. Bibliographical References

Bosque-Gil, J., Mititelu, V. B., Oliveira, H. G., Ionov, M., Gracia, J., Rychkova, L., Oleskeviciene, G. V., Chiarcos, C., Declerck, T., and Dojchinovsk, M. (2021). Balancing the digital presence of languages in and for technological development. A Policy Brief on the Inclusion of Data of Under-resourced Languages into the Linked Data Cloud. https://nexuslinguarum.eu/results/policy-briefs.

Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). *Linguistic Linked Data: Representation, Generation and Applications*. Springer International Publishing.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G.-B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., et al. (2022 (in press)). When linguistics meets web technologies. recent advances in modelling linguistic linked open data. *Semantic Web Journal*.

Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., Van Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., et al. (2020). Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. *Semantic Web*, 11(5):855–881.