Guidelines and Best Practices on Linguistic Linked Open Data

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 Authors	Patricia Martín Chozas, Milan Dojchinovski, Katerina Gkirtzou, Anas Fahad Khan, Andon Tchechmedjiev
Contributors	W3C BPMLOD members
Version | Status	v1.0 | final
Date	13 February 2024

Acronyms List

CA	COST Action
LD	Linked Data
LLD	Linguistic Linked Data
LLOD	Linguistic Linked Open Data
LOD	Linked Open Data
NLP	Natural Language Processing
WG	Working Group

List of Tables

Table of Contents

EXECUTIVE SUMMARY

The document is about the collaborative efforts between the Nexus Linguarum COST Action and the Best Practices for Multilingual Linked Open Data (BPMLOD) W3C Community Group to develop guidelines and best practices for linking data and services across languages. It discusses the creation of two sets of guidelines: one focusing on Linked Linguistic Open Data (LLOD) and the other on integrating LLOD with Natural Language Processing (NLP) services. The text summarizes the work done by Working Group 1 (WG1) and Working Group 2 (WG2) within NexusLinguarum, highlighting their contributions to the development of comprehensive recommendations for LLOD and LLOD-aware NLP services. Additionally, it discusses the evolution of the BPMLOD community, ongoing efforts to update guidelines, and plans for sustainability beyond the Nexus Linguarum COST Action.

1. Introduction

One of the main goals of the Nexus Linguarum COST Action is "to propose, agree upon and disseminate best practices and standards for linking data and services across languages."[1] This has led to the development of two sets of guidelines and best practices, the first dealing with the interlinking, publication, and validation of LLOD and the second with guidelines and best practices for LLOD and NLP. The first set of guidelines and best practices have been developed within the scope of the Nexus Linguarm Working Group 1 (WG1) , while the second set of guidelines and best practices have been developed as part of Nexus Linguarum Working Group 2 (WG2). In this deliverable we look at both sets of guidelines and best practices. The chairs of the WG1 and WG2 had originally intended to produce two separate deliverables reporting on each. However as work proceeded in WG's 1 and 2 the overlap between the two sets of guidelines was felt to be strong enough that a shared means of organising the work in both was proposed via the BPMLOD group (see Section 2 below) and eventually it was decided that a shared deliverable would be the most suitable option. However, an account of the background and preparatory work carried out in WG1 can be found in Section 3.1, and that carried out in WG2 in Section 3.2.

2. The Best Practices for Multilingual Linked Open Data W3C Community Group

The Best Practices for Multilingual Linked Open Data (BPMLOD) W3C Community Group[2] was formed in 2013 out of the need of the multilingual web community to establish best practices for the creation and linking of Linked (Open) Data in multiple languages. The BPMLOD community operates on a bottom-up approach, emphasising community-driven efforts and close collaboration to address the challenges and complexities associated with multilingual linked data. In 2015, much prior to the start of Nexus Linguarum, the BPMLOD community published in total eight guideline community reports, six related to WG1 tasks and objectives, one regarding Linguistic Linked Data exploitation (WG4), and two related to WG2 tasks and objectives.

Over the past ten years, the multilingual web has continuously evolved offering new resources as well as new vocabularies and ontologies for the creation of Multilingual Linguistic Linked Open Data. The main driving forces behind this growth consist of the Ontology-Lexica Community Group W3C Community Group[3] and projects, like the NexusLinguarum COST Action. These have created the necessity to update existing Guidelines and Best Practices for Multilingual Linguistic Linked Data, as well as to explore new topics in light of the experience of the last ten years or so, since the initial published reports of the BPMLOD community. As a result, the BPMLOD community has worked in close relation with the NexusLinguarum COST Action in pursuit of their shared aims. At the moment, the BPMLOD community consists of 93 participants with numerous new participants coming from NexusLinguarum. The BPMLOD community holds a number of calls: one per topic of interest (for details of each topic see Section 3) and a regular central call convened by the BPMLOD chairs for the coordination of the whole community across all topics. These calls are announced on the BPMLOD public mailing list and the BPMLOD wiki page.

[1] See the Nexus Linguarum memorandum of understanding: https://e-services.cost.eu/files/domain_files/CA/Action_CA18209/mou/CA18209-e.pdf
[2] https://www.w3.org/community/bpmlod/
[3] https://www.w3.org/community/ontolex/

3. Guidelines and best practices on LLOD

This section summarises the work on guidelines and best practices on LLOD within two task forces:

1) Guidelines and best practices on generation, interlinking, publication, and validation of LLOD, discussed in Section 3.1, and
2) Guidelines and best practices on LLOD and NLP discussed in Section 3.2.

In the table below, Table 1, we list the topics related to the guidelines and best practices of both task forces, along with their scope and their current status.

Topic	Scope	Status[4]
Terminology generation (TBX)	Best practices for transforming multilingual terminologies, particularly those available in TBX format, into a Linked Data version.	Initiated, early stage.
Bilingual dictionaries	Aimed to guide in the process of creating a linked data (LD) version of a lexical resource, particularly a bilingual dictionary.	Draft report completed, waiting for publication.[5]
Crosslingual linking[6]	Developing BPMLOD guidelines, best practices, use cases for cross-lingual interlinking, related to NexusLinguarum T1.3 activities.	Regular calls, first draft under preparation.
Corpora annotation (NIF, Web annotation)	Best practices to follow for the generation of Linked Data text corpora, using the NLP Interchange Format (NIF).	Initiated, organised a few calls. Checked the state of the art. Low interest from the community.
Guidelines for Developing NIF-based NLP Services	Best practices to follow for the implementation of RESTful NLP web services that rely on the NLP Interchange Format (NIF).	Initiated, organised a few calls. Checked the state of the art. Low interest from the community.
Wordnets	Best practices on converting and modelling wordnets to Linked Data.	After an initial call we decided not to update the BPMLOD guidelines on the conversion and modelling of wordnets in RDF since the Global WordNet Association website/github[7] has enough

[4] As of 31 Jan 2024
[5] Draft report for Bilingual dictionaries: https://bpmlod.github.io/Bilingual_Dictionaries_Report/
[6] Git repository with material for Crosslingual Linking: https://github.com/bpmlod/Crosslingual_linking_report
[7] https://globalwordnet.github.io/schemas/

		guidance on this task but in future we plan to produce guidelines for more specific kinds of WordNet.
LLOD for Under-resourced languages	Best practices for under-resourced languages and LLOD.	Initiated, initial document prepared, bibliography collected.
Neuro-symbolic LLOD[8]	Best practices for Neuro-symbolic LLOD.	Regular calls, gathering material on both data and services aspects
Licensing and metadata[9]	Guidelines for licensing language resources.	Regular calls, first draft under preparation.
~~***Multimodal data***~~	Guidelines on modelling multimodal information as Linked Data.	Decision to drop this topic as there is already an ongoing effort in that direction.
Sign languages	Guidelines on modelling sign language information as Linked Data.	Initiated, regular calls to identify existing efforts and material. Surveying what exists and drawing perspectives for future directions
Lexicographic LLOD[10]	Guidelines Lexicographic LLOD.	Initiated, drafting guidelines started, still early draft.
Encoding Domain Labels for Linked Data Lexical Resources in RDF	A series of guidelines for encoding domain label information in RDF using three linked data vocabularies, namely OntoLex-Lemon, SKOS, and lexicog.	First draft completed[11]

Table 1: List of Guidelines and Best Practices proposed so far

[8] Git repository with material for Neurosymbolic LLOD: https://github.com/bpmlod/Neuro-symbolic_LLOD_report
[9] Git repository with material for Licensing and metadata: https://github.com/bpmlod/Licensing_and_metadata_report
[10] Git repository with material for Lexicographic LLOD: https://github.com/bpmlod/Bilingual_Dictionaries_Report
[11] Note that in developing this set of guidelines the authors of the first draft were able to make use of a short term scientific grant provided by Nexus Linguarum, see https://nexuslinguarum.eu/blog-post-on-the-stsm-domain-labels-in-linked-lexicographic-resources-by-fahad-khan-at-lisbon-portugal/ . The first draft can be found here https://github.com/anasfkhan81/EncodingDomainLabelsRDF/blob/main/Guidelines.md

3.1 Guidelines and best practices on generation, interlinking, publication and validation of LLOD

Prior to initiating the work which is described below the members of WG1 carried out a survey of already existing materials. The details of this survey are given in Khan et al 2022; here we will give a brief summary of our findings focusing on materials which were primarily intended to fulfil the role of Guidelines and/or Best Practices for the tasks of generation, interlinking, publication, and validation of LLOD. Our main finding concerned the dearth of suitable materials answering to this description.

Those few sets of materials which we did discover tuned out fit into two categories, the first of which were sets of best practices produced in the past by the BPMLOD W3C community group in its first incarnation (see above, Section 2), a total of eight sets of guidelines, and the second of which were an output of the LIDER project[12], a series of eight reference cards. The former comprise guidelines for generating multilingual[13] and bilingual[14] dictionaries, wordnets[15], TBX terminologies[16]; as well as for developing NIF services[17] and LLOD aware services[18] and creating corpora with NIF. Finally, the BPMLOD group also published guidelines for LLD exploitation[19]. All of the preceding documentation was published in 2015, almost a decade from the time of writing, and a year before the latest version of *lemon*, that is, OntoLex-Lemon, was published[20]. Since there are numerous classes and properties which exist in OntoLex-Lemon and are not yet in the preceding version of *lemon*, and vice versa, those guidelines which reference this vocabulary clearly stand in need of updating. In addition, the 2015 BPMLOD guidelines do not take *lexicog* into consideration, that is, the extension of the original OntoLex-Lemon vocabulary which deals specifically with lexicographic resources[21], since this was, once again, published subsequently (in 2019). However, this new vocabulary clearly has an impact on the first two dictionary related BPMLOD guidelines. As well as being out of date, the original BPMLOD guidelines do not cover all tasks and all types of resources.

Moving onto the eight reference cards which were made available by the **LIDER project[22]**. These consist of short guides to **publishing linked data**[23], **language resource licensing**[24], **inclusion in the LLOD cloud**[25], **data**

[12] https://lider-project.eu
[13] http://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/
[14] https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/
[15] http://bpmlod.github.io/report/WordNets/index.html (Unofficial Draft)
[16] https://www.w3.org/2015/09/bpmlod-reports/multilingual-terminologies/
[17] https://www.w3.org/2015/09/bpmlod-reports/nif-based-nlp-webservices/
[18] http://bpmlod.github.io/report/LLOD-aware-services/index.html
[19] https://www.w3.org/2015/09/bpmlod-reports/lld-exploitation/
[20] https://www.w3.org/2016/05/ontolex/
[21] https://www.w3.org/2019/09/lexicog/
[22] https://cordis.europa.eu/project/id/610782
[23] http://bpmlod.github.io/report/LLOD-aware-services/index.html
[24] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/How-to-publish-linguistic-linked-data-Reference-Card.pdf
[25] https://lider-project.eu/lider-project.eu/sites/default/files/referencecards/Inclusion-in-the-LLOD-Cloud-Reference-Card.pdf

IDs[26], **language resource discovery with Linghub**[27], **NIF corpora**[28], the representation of **crosslingual links**[29] and **language resource documentation in datahub**[30]. Each card features 'how to' instructions addressing different target groups, with a clear description of the steps and the knowledge needed to carry out a particular task or set of tasks along with whatever resources or tools useful for reaching the goal. All of the cards date from 2015 (an exemplary year for LLD guidelines and best practices it seems!). As far as we have been able to ascertain, there is no licensing information available for these reference cards making it difficult to re-use the materials they contain. Since the reference cards run to two pages each they are limited in the amount of information they offer for any given task or goal.

In order to address the above-mentioned issues, members of the NexusLinguarum COST Action decided to initiate new efforts within the scope of the BMPLOD W3C community group[31], see Section 2. After a process of consultation with the NexusLinguarum community, 12 potential topics were identified for a new set of Guidelines and Best Practices related to the Working Group 1 task force. For more details, about the topics and status see Table 1 from Section 3.

3.2. Guidelines and best practices on LLOD and NLP

Working Group 2 is responsible for leveraging Language Resources published following the Linked Data principles to enhance various Natural Language Processing tasks, such as Knowledge Extraction, Machine Translation, Question Answering, and more. As a result, within WG2, a series of synergies have emerged focused on research of techniques that may benefit from the semantics of LLOD, such as knowledge enhanced neural networks or exploiting Linked Data for Entity Linking and Word Sense Disambiguation amongst other tasks.

The main initiative derived from such synergies is a group-level effort to review literature describing LLOD-aware NLP approaches. This effort involves around 20 participants from WG2 with a very varied background and it is lead by Andon Tchechmedjiev, who has taken care of the automatic collection and clustering of the papers and their assignment to participants in the survey and the creation of a set of paper screening guidelines to standardise the annotation process.

---

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

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

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

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

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

[31] See https://www.w3.org/community/bpmlod/wiki/Guidelines_and_best_practices_for_LLOD

We consider this literature review a good starting point to discover guidelines and best practices for the exploitation of LLOD by NLP services. Regarding the process of the survey, we are still finishing the second annotation phase as per the following criteria:

- We understand LD-aware NLP as:
    - NLP that makes use of LD as an integral part of it
    - NLP that is ready to use LD and is explicitly designed to do so to take advantage of LD capabilities such as dynamicity, interoperability, reasoning, etc.
    - NLP that uses LD to describe or specify its pipelines, tasks, or processes to facilitate interoperability
- We do not understand LD-aware NLP as:
    - NLP that use or could use LD as a source of data merely, not being aware of the characteristics of LD
    - NLP that is used to generate LD but without LD being relevant to the NLP algorithm(s)

Papers within the first group are selected to be included in the survey paper. Examples of the content of such papers are:

- An Information Extraction tool that uses a LD dataset as a source of entities and relations but also exploits the links to other entities.
- An NLP system that uses an ontology to specify its pipeline/workflows.
- A Machine Learning tool that exploits RDF bilingual dictionaries taking advantage of their links to infer new possible translations.
- A KG embedding method that uses Web-scale or cloud-scale federated learning on KGs, exploiting LD properties/capabilities (e.g. query federation) within the approach

The data extraction phase of the survey will also allow tagging papers that contain relevant information to derive recommendations, which will allow a first coordination meeting to be held regarding LD-aware NLP services. A closely related recommendation effort that is underway pertains to NIF Web Services, which often support LD-aware NLP approaches, but using a specific standardised format and associated tooling. Both efforts are aimed at updating and extending the already existent BPMLOD guidelines on developing NIF services and LLOD aware services referred in Section 3.1

Beyond LD-aware NLP services, another particularly popular aspect pertaining to LD-awareness is the consideration of neuro symbolic approaches integrating LD with deep learning architectures. Although methodologically more under the purview of WG3, the recommendation efforts on NeuroSymbolic learning within BPMLOD, equally involve participants from both working groups WG1 and WG2.

4. Sustainability and Future Directions

As NexusLinguarum comes to an end, 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 wonderful tools provided

by COST. As described throughout this deliverable, the cornerstone of our sustainability strategy lies in establishing or reviving community groups and efforts that would pursue the efforts durably in time.

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. We submitted a COST Innovator Grant, with a focus on increasingly involving industrial actors and creating the foundation for long-term financing. 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.

5. Conclusions

The Nexus Linguarum COST Action, alongside the Best Practices for Multilingual Linked Open Data (BPMLOD) W3C Community Group, has significantly contributed to the development and dissemination of guidelines and best practices for linking data and services across languages. The collaborative efforts of Working Group 1 (WG1) and Working Group 2 (WG2) within NexusLinguarum have resulted in comprehensive recommendations spanning the generation, interlinking, publication, and validation of Linked Linguistic Open Data (LLOD) as well as its integration with Natural Language Processing (NLP) services.

The BPMLOD community has evolved, producing several guideline reports and fostering collaborations with initiatives like NexusLinguarum. However, the need for updating existing guidelines and exploring new topics has been recognised, leading to ongoing efforts within the BPMLOD community.

WG1's focus on LLOD has revealed gaps in existing materials, prompting initiatives to develop new guidelines and best practices. Similarly, WG2's emphasis on LLOD-aware NLP services highlights the importance of leveraging semantic data for enhancing various NLP tasks. The ongoing literature review and recommendation efforts within WG2 demonstrate a commitment to advancing the integration of LLOD into NLP methodologies.

Looking ahead, sustainability is a key consideration for the continuation of these efforts beyond the NexusLinguarum COST Action. Community-led initiatives, such as the BPMLOD and Ontolex-Lemon W3C community groups, are poised to sustain the development and adoption of guidelines and standards. Efforts to secure long-term financing through avenues like the COST Innovator Grant and future project proposals underscore a commitment to ensuring the enduring impact of these initiatives.

In conclusion, the collaborative endeavours of NexusLinguarum and BPMLOD have significantly advanced the field of multilingual linked data and NLP, laying a solid foundation for future developments and collaborations within the broader Linguistic and Semantic Web community.

Bibliography

- Fahad Khan, Christian Chiarcos, Thierry Declerck, Maria Pia Di Buono, Milan Dojchinovski, Jorge Gracia, Giedre Valunaite Oleskeviciene, and Daniela Gifu. 2022. A Survey of Guidelines and Best Practices for the Generation, Interlinking, Publication, and Validation of Linguistic Linked Data. In Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, pages 69–77, Marseille, France. European Language Resources Association.