Nexus Linguarum

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

# D1.5

# Final Activity Report

# Working Group 1 “Linked data-based language resources”

**Main authors:**

Milan Dojchinovski and Anas Fahad Khan

| | |
|---|---|
| **Project Acronym** | NexusLinguarum |
| **Project Title** | European network for Web-centred linguistic data science |
| **COST Action** | 18209 |
| **Starting Date** | 27 April 2024 |
| **Duration** | 54 months |
| **Project Website** | https://nexuslinguarum.eu/ |
| **Chair** | Jorge Gracia |
| **Main authors** | Milan Dojchinovski and Anas Fahad Khan |
| **Contributors** | Verginica Barbu Mititelu, Maria Pia di Buono, Maxim Ionov, David Lindemann, Mike Rosner, Blerina Spahiu, Ranka Stanković, Vojtěch Svátek |
| **Reviewer** | NexusLinguarum core group team |
| **Version \| Status** | final |
| **Date** | 26/04/2024 |

## Acronyms List

| | |
|---|---|
| CA | COST Action |
| ISO | International Organization for Standardization |
| LMF | Lexical Markup Framework |
| LD | Linked Data |
| LD4LT | Linked Data for Language Technology |
| LLD | Linguistic Linked Data |
| LLOD | Linguistic Linked Open Data |
| LOD | Linked Open Data |
| LR | Language Resource |
| NLP | Natural Language Processing |
| RDF | Resource Description Framework |
| SOTA | State Of The Art |
| STSM | Short Term Scientific Mission |
| TEI | Text Encoding Initiative |
| UC | Use Case |
| WG | Working Group |

# Table of Contents

# Executive Summary

This report summarises the progress and the status of the work of Working Group 1 (WG1), “Linked data-based language resources”, as part of the NexusLinguarum COST Action (CA) CA18209 during its second Action period, i.e. month 25 until month 54. Over the course of these two and half years WG1 has successfully managed to further advance the tasks activities and organise a number of events and meetings. The WG has fulfilled its defined goals for the second half of the Action, namely: starting to establish foundations for development, publishing, modelling, linking, enrichment, quality assurance and repair of Linguistic LOD resources. The WG has addressed these goals by the active collaboration of its members, the organisation of events such as training schools and workshops, the execution of Short Time Scientific Missions and Virtual Mobility Grants, and the dissemination activities.

# 1. Introduction

In this final activity report for Working Group 1 (WG1) of the Nexus Linguarum COST action we summarise the activities carried out within the ambit of that working group in the second period of the Action. In what follows in this introduction we will give a brief summary of WG1 and the work which it has been carrying out during the COST action.

Language resources (LR) play a key role in research in humanities and in the development of NLP applications. The aim of **WG 1** was to make the benefits of the linked data paradigm accessible to creators, managers and consumers of LRs in order to make them more easily discoverable, reusable and interoperable both with one another and with the tools working with them. In pursuit of this aim, the group addressed the creation, interlinking, enrichment, quality assessment and evolution of LR from a linked data perspective.

Throughout the life cycle of a language resource, it is vital to take into account the features of the language(s) concerned to identify which aspects are language-dependent, and thus provide equal support to both resourced and under-resourced languages. As one of the goals of our Action was to help bridge the wide gap in technological support and available data between resourced and under-resourced languages, in WG1 we had a task force (**Task 1.5**) concerned specifically with the analysis and development of language technologies for under-resourced languages and domains from a linked data perspective.

As a starting point in the creation of a linked-data based language resource users are required to analyse which features are present in the original data, assess their representation needs, and select, adapt or create a linked data vocabulary to reuse in order to represent the encoded information as linked data. This step is focused thus on the modelling of linguistic information as linked data which was the topic of **Task 1.1**.

Creating a linked data-based language resource however concerns more than just modelling. To contribute to the development of language resources in a collaborative setting, WG1 also worked on analysing the current landscape of linked data resources published on the LLOD cloud, taking into account their availability, their quality, and the languages and domains they cover. This allowed us to gather insights into the tendencies in language resource publication, the use of metadata, and gaps in language and domain coverage needing solution. In addition, having undertaken to promote and support the adoption of LD principles in the creation of language resources, we also investigated the main obstacles preventing their (re)use, together with the need for (semi)automatic tools/services to enhance the exploitation of linked data. In particular, **Task 1.2** dealt with the creation and evolution of LLOD resources in a distributed and collaborative setting, **Task 1.3** with cross-lingual data interlinking, access and retrieval in the LLOD) and **Task 1.4** with improving and monitoring quality of LLOD sources.

For the second half of the NexusLinguarum CA, WG1 has contributed to the following set of deliverables:

D1.4 Guidelines and Best Practices on Linguistic Linked Open Data

D3.2 Report and Training Materials of the 3rd Training School

D1.5 Final Activity Report Working Group 1 "Linked data-based language resources" (this document)

The final composition of leadership in Working Group one is presented in the following table.

| Role | Person | Country |
|---|---|---|
| WG1 leader | Milan Dojchinovski | Czech Republic |
| WG1 co-leader | Anas Fahad Khan | Italy |
| Task 1.1 leader and co-leader | Christian Chiarcos | Germany |
| | Anas Fahad Khan | Italy |
| Task 1.2 leader and co-leader | Maria Pia di Buono | Italy |
| | Verginica Mititelu | Romania |
| Task 1.3 leader | Mike Rosner | Malta |
| Task 1.4 leader and co-leader | Blerina Spahiu | Italy |
| | Vojtech Svatek | Czech Republic |
| Task 1.5 leader and co-leaders | Max Ionov | Germany |
| | David Lindemann | Spain |
| | Ranka Stankovic | Serbia |

**Table 1.** Structure of WG1 (as of April 15th, 2024).

The remainder of this report is as follows: **Section 2** provides detailed information about the progress and the status for each of the WG1 tasks. **Section 3** provides a brief overview of the guidelines and best practices efforts. **Section 4** provides a summary of the organised WG1 related events. **Section 5** outlines the future directions of the WG1 and provides an overall summary of the report. Finally, **Section 6** lists the outcome publications generated by WG1 members in the context of the different tasks of the WG.

# 2. Tasks Reports

## 2.1. Task 1.1 LLOD modelling

**Task Leaders:**

- Leader: Christian Chiarcos
- Co-leader: Fahad Khan

**General Overview**

In the period covered by the final report, as throughout the whole action, Task 1.1 focused on the analysis of modelling needs of language resources and the development of best practices in collaboration with other standardisation initiatives. In particular it dealt with the updating, extension and improvement of currently existing models for representing linguistic information as linked data (Ontolex-lemon, OLiA, NIF, etc.). It collaborated with the W3C Ontology Lexicon group on the development of two OntoLex-Lemon modules and contributed to discussions on two others (see Progress below for more details). As well as this it continued work on the alignment of different standards for linguistic annotation.

**Progress as of M54**

- Close collaboration with the W3C Ontology Lexicon Group on the development of new Ontolex modules, or the evaluation of whether proposed modules are needed: Thanks to this collaboration, two new OntoLex modules FRequency And Corpus (FrAC) and Morph are due to be published soon.
- Task 1.1 helped (together with Task 2.5) in setting up new Terminology module discussions as well as participating in regular discussions on a prospective new multimodality and multimedia module, led by Thierry Declerck (RIP).
- Continued collaboration with the W3C Linked Data for Language Technology (LD4LT) group on alignment of different standards for linguistic annotation (including contributions to LDK2023 W3C days and [planned] to LDL2024).
- COST action collaboration with WG4 Use case UC4.2.1.
- Several articles have come out of Task 1.1 discussions including ‘When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data’, 2021.
- Task 1.1. had one dedicated STSM: “Domain Labels in Linked Lexicographic Resources” by Fahad Khan who was hosted at the Universidade Nova in Lisbon, Portugal.

**Future plans**

Future plans include:

- Publication of FrAC an Morph modules.
- Continuing collaboration on Terminology module via the W3C Ontology-Lexicon group.

## 2.2. Task 1.2 Creation and evolution of LLOD resources in a distributed and collaborative setting

**Task Leaders:**

- Leader: Maria Pia di Buono
- Co-leader: Verginica Barbu Mititelu

**General Overview**

The goal of this task is to describe the creation and evolution of LLOD resources in a distributed and collaborative setting. The task activities have been focussed on three main aspects: (i) repositories' coverage and accessibility/availability of LD resources; (ii) adoption of LD principles among different types of users; (iii) (semi) automatica generation of LD resources. Starting from an evaluation of the state-of-the-art of linked data (LD) resources conducted together with Task 1.4 and Task 1.5, we propose a metadata enrichment to improve findability and accessibility of LD resources. A further analysis has been conducted to identify the potential obstacles preventing LD reuse, extension, and creation. Finally, in the last part of the Action, the focus has been moved to a preliminary evaluation of large language models (LLMs) in supporting the creation of LD resources. The results of the activities are further detailed below.

**Progress as of M54**

- With reference to the first topic, in collaboration with Task 1.4 and Task 1.5, an extensive analysis of the metadata of the LD resources in two repositories (LOD Cloud[1] and Annohub[2]) was carried out. Although all metadata fields are important, we particularly focused on three of them: language, domain and type, the last two especially with regard to linguistically relevant resources. Their investigation revealed the need to deal with inconsistencies and gaps in information representation, all of which meant some automatic, but mainly manual intervention in the respective fields. The analysis offers insights into the two repositories' coverage of LD resources for different languages, of different types and for different domains. Furthermore, we evaluate the accessibility/availability of existing linguistic LD resources (see Task 1.4).

  The description of the whole process, the decisions made and the results obtained are all presented in the paper "Paving the Way for Enriched Metadata of Linguistic Linked Data", published in a special issue of the Semantic Web Journal on Linguistic Linked Data and still under review. Proposal of META-SHARE Enriched LLD (MELLD), enriched and META-SHARE aligned metadata for the resources in the two repositories, made available in GitHub[3].

- In order to evaluate the potential obstacles to the adoption of LD principles, a survey on the obstacles for LD reuse, extension, and creation[4] has been conducted. Designing this survey meant to offer a clearer idea on what prevents people from:
  - (re)using existing resources,
  - extending such resources,
  - creating new LD resources,
  - using one of the available vocabularies/schemas.

  The survey was sent out to various mailing lists. The (preliminary) results were presented in the WG1 meeting in the NexusLinguarum MC meeting (29th September). Given that we got only 43 responses during the first call (summer 2021)[5],

[1] https://lod-cloud.net
[2] https://annohub.linguistik.de/
[3] https://github.com/unior-nlp-research-group/melld
[4] https://forms.gle/aaCHV1fsxM9CbJjA7
[5] https://docs.google.com/spreadsheets/d/102XbYrcVnw-A4bo-D7MhorPX4WCKnohHYSHKZlHMwR8/

we have made the decision to rethink the strategy of attracting more participants.
The final results of the survey have been accepted at LDK 2023 conference and awarded with the Best Poster presentation (ex aequo).

- The evaluation of LLMs to support the creation/conversion of/into linguistic linked data (LLD) has been conducted on a small set of data for four languages that are English, Albanian, Italian and Romanian. The study investigates the potential of LLMs for knowledge formalisation using well-defined vocabularies, specifically focusing on OntoLex-Lemon. The research questions addressed are the following: (i) How will this new paradigm of human-machine interaction impact established knowledge representation formalisms?; (ii) Are LLMs ready to contribute to knowledge formalisation using well-defined ontologies? (iii) Do these models perform consistently across different languages?
  The initial experiment evaluates GPT 3.5 in the generation of RDF resources using two different types of prompting, i.e., zero-shot and few-shot prompts. The results have been analysed through a multidimensional approach, considering both the general outputs and the RDF outputs.
  Despite the limitations, the application of LLMs for generating LLD seems quite promising under the assumption of adopting specific strategies of prompting to ensure the result robustness. In the future, we plan to implement a post-generation filtering system that performs some sanity checks and adaptive prompting to improve the quality of the LLM output by identifying and correcting errors, leading to more reliable results. The results of this experiment have been described in a paper accepted at the workshop on Deep Learning and Linked Data (DLnLD) - LREC-COLING 2024.

*Synergies with other tasks:* during the action we have been collaborating closely with the leaders and co-leaders of Task 1.5 and Task 1.4, as investigation of the languages represented in the repositories naturally leads to the discovery of under-resourced ones, while metadata assessment also implied checking the availability of the resources dump files and/or SPARQL endpoint, given that lack of maintenance is one of the frequent complaints with respect to language resources in general.

**Future plans**

- Further evaluating LLMs to support the creation and conversion of LD resources
- Proposing a short users' guide on LLM prompting to generate LD resources
- Publishing the results of the experiments.

## 2.3. Task 1.3 Cross-lingual data interlinking, access and retrieval in the LLOD

**Task Leaders:**

- Leader: Mike Rosner[6]

[6] Sina Ahmadi was acting as co-leader during the period 2021-22.

**General Overview**

Work on this task began by carefully analysing the requirements given in the MOU, according to which the focus is on three aspects: (i) cross-lingual data interlinking, access and retrieval in the LLOD; (ii) methods and techniques based on LLOD for accessing and exploiting data across different languages and (iii) novel (semi-) automatic methods that aim at increasing the amount and level of interlinking across LLOD datasets. These three foci formed the basis of research. The subgroup has carried out its work through a series of regular telcos since 2021.

During the period 2021-22 our efforts were mostly focused on (i) cross-lingual link discovery, addressing in particular the potential for LLOD to contribute to resource development of under-resourced languages by discovering cross-lingual links automatically. This culminated in a publication presented at LREC 2022. Subsequently attention shifted towards aspect (ii) with an emphasis on multilingualism leading to collaborations with several other task groups, notably tasks 1.1 , 1.5, 3.1, 3.3. Consideration of the multidimensionality aspect of LLOD led to ongoing contributions to Task 3.4. These joint efforts are further detailed below.

**Progress as of M54**

- Collaboration with task 1.1 (LLOD Modelling) and 1.5 (Development of the LLOD cloud for under-resourced languages and domains). The modelling task T1.1 was concerned with finalising the specification of Ontolex Morph. We developed a working demonstration than Ontolex Morph is sufficient for handling Maltese, an under-resourced Semitic Language. This resulted in a paper presented at the LDK 2023 Workshop held in Vienna. A second paper which examines and exemplifies the concept of interoperability using LLOD for morphological description has been accepted for presentation at the LDL Workshop, LREC2024.
- Collaboration with task 3.3 (Linking structured multilingual language data across linguistic description levels) began by investigating the role of LLOD in linking multilingual data across different linguistic description levels. Since M54 this work has been distilled into an accepted SWJ article to appear during 2024.
- Collaboration with task 3.1 (Big Data and Linguistic Information) explored the opportunities for mutual enhancement between the techniques of LLOD and and those of Big Data. Our ideas were developed into a position paper which was accepted for presentation at LREC 2024
- Collaboration with task 3.4 (Multidimensional linguistic data) began by surveying the current state of the LLOD cloud with respect to multimodal data. However, this developed into an investigation into the use of LLOD for sign-language, a paradigm exemplar of both multimodality and multilinguality, since different sign languages are indeed distinct languages with their own vocabulary. Work on this topic is ongoing. A preliminary document provides an overview of the work on this subject. The intention for this to be completed and published as a state-of-the-art survey.
- In the second quarter of 2023 T1.3 participants began to focus on developing BPMLOD guidelines for cross-lingual interlinking, since there are many ways in which such interlinking can be implemented but little guidance on how to choose between

the available possibilities. Two documents are planned: a first version of the guidelines to be published under the auspices of the BPMLOD group of the W3C, for which a preliminary draft is already available. A second, more theoretical position paper on the nature and scope of cross-lingual links is planned for submission to NLE.

**Future plans**

- Finalise W3C guidelines on cross-lingual interlinking.
- NLE journal article on cross-lingual interlinking levels.
- Publish survey on LLOD and sign-language.

## 2.4. Task 1.4 Improving and monitoring quality of LLOD sources

**Task Leaders:**

- Leader: Blerina Spahiu
- Co-leader: Vojtech Svatek

**General Overview**

This task has the aim to monitor and improve the quality of LLOD sources by novel metrics and approaches to diagnosis and repair. The activities within this task are focused on analysing and developing semi-automatic and automatic methods for validating linked data and cross-resource links via collaborative strategies.

**Progress as of M54**

- We conducted a comprehensive analysis of the LLOD cloud datasets to identify key structural characteristics of such knowledge graphs. To achieve this objective, we proposed a set of specific metrics specifically designed to capture these structural features. Our analysis incorporated over 20 new metrics into the computed profile, including: Number of entities Number of triples Average number of triples per entity Internal and external concepts and properties relative to the ontology These metrics were applied to analyse linguistic data, and the resulting insights provide a foundation for a more efficient understanding of linguistic information within knowledge graphs.

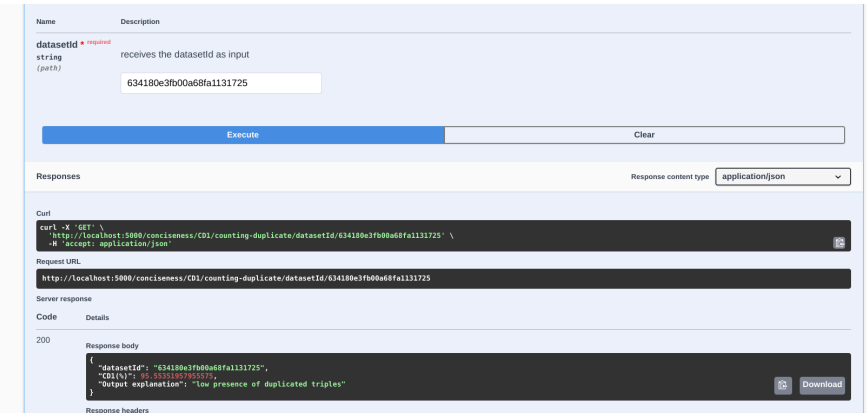
  Such analysis is published in “Blerina Spahiu, Renzo Alva Principe, and Andrea Maurino. Profiling Linguistic Knowledge Graphs. *Proceedings of the 4th Conference on Language, Data and Knowledge*. 2023”.

- We investigated the utilisation of machine learning methods for pruning and re-ranking frequent patterns in knowledge graph profiling to enhance semantic understanding. We manually labelled entities to create an input dataset and used various machine learning algorithms like random forest and logistic regression for pattern post-processing. Our findings suggest that machine learning techniques can enhance the efficiency of knowledge graph profiling by improving pattern prioritisation.

  Such study is published in “Gollam Rabby, Farhana Keya, Vojtech Svátek, Blerina Spahiu - Pruning and re-ranking the frequent patterns in knowledge graph profiling

using machine learning. *Proceedings of the 4th Conference on Language, Data and Knowledge. 2023."*

- We evaluated the quality of LLOD resources. For this aim we have implemented 29 quality metrics among which: consistency, availability, licensing, interlinking, provenance, syntactic validity, etc. Such metrics are implemented as REST API requests. Below there is a screenshot of the concisiness dimension, for the metric duplicated triples. These metrics are then applied on the LLOD cloud datasets. This study is under preparation for a submission to a journal.




- As in Section 2.2, in collaboration with the task T1.2 we conducted a survey on users' experiences with LLD principles. We aimed to assess the impact, challenges, and opportunities in adopting LLD. Through the survey, we gathered information on participants' backgrounds, LLD knowledge, usage, development, publishing, and metadata utilisation. Our study provides valuable insights for enhancing the adoption of LLD principles across different user groups.

  This study has been awarded with the Best Poster at LDK and is published in "Verginica Mititelu, Maria Pia Di Buono, Hugo Gonçalo Oliveira, Blerina Spahiu, Giedrė Valūnaitė-Oleškevičienė - Adopting Linguistic Linked Data Principles: Insights on Users' Experience.- *Proceedings of the 4th Conference on Language, Data and Knowledge. 2023."*

- As in Section 2.2, in collaboration with the task T1.2, we explored the potential of large language models (LLMs) for knowledge formalisation using well-defined vocabularies, focusing on OntoLex-Lemon. The study tests LLMs for four languages (English, Italian, Albanian, Romanian) and analyses the formalisation quality of nine words with varying characteristics. The research aims to initiate a discussion on the potential and challenges of utilising LLMs for knowledge formalisation within the Semantic Web framework.

  This study is published on "Maria Pia Di Buono, Blerina Spahiu, Verginica Mititelu - Evaluating Large Language Models for Linguistic Linked Data Generation - In *Proceedings of the First Workshop DLnLD: Deep Learning and Linked Data-LREC-COLING 2024-The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation."*

- Within this task we also initiated and organised two editions of the “Profiling Linguistics Knowledge Graphs” (ProLingKNOWER).
    - The 1st edition of the ProLingKNOWER workshop held on 23 May 2022, co-located with NexusDays in Jerusalem, Israel.
    - The 2nd edition of the ProLingKNOWER workshop held on 12 September 2023, co-located with LDK 2023 in Vienna, Austria.

**Future plans**

- Publish the result of the quality evaluation of LLOD: We intend to publish the results from the quality evaluation analysis of LLOD datasets. This publication is under preparation.
- Develop an interactive interface: We intend to develop an interactive interface where users might profile and quality evaluate their KG regardless of the domain. This interface will allow comparison between different features of different datasets.
- Keep engaging with fellow members of working groups within NexusLinguarum.
- Uphold the tradition of organising the ProLingKNOWER workshop.

## 2.5. Task 1.5 Development of the LLOD cloud for under-resourced languages and domains

**Task Leaders:**

- Leader: Maxim Ionov
- Co-leader: David Linderman
- Co-leader: Ranka Stankovic

**General Overview**

In the period covered by the deliverable this task focused on the development of several wikibases, a literary corpus in NIF, several datasets in OntoLex and OntoLex-Morph.

The free **Wikibase** software enables instant publishing in accordance with the Semantic Web and FAIR principles, and the collaborative curation of datasets. Wikibase (https://wikiba.se) as an extension of MediaWiki is the software underlying Wikidata (https://www.wikidata.org) (Vrandečić and Krötzsch, 2014), a very large knowledge graph maintained by the community of Wikidata users, and technically supported by Wikimedia Deutschland (WMDE, https://www.wikimedia.de). Wikibase can be used for creating data archives that can easily interact with the semantic web through the use of open standards; compared with other software solutions, it offers unique features, such as the option to manually edit every single semantic triple in a graphical interface, user and user rights management, reversible edit histories, a graphical SPARQL query interface and a programmatically queriable endpoint, and an API, among those of any usual MediaWiki instance. Wikibase Cloud (https://www.wikibase.cloud/) is a free Wikibase hosting service provided by WMDE.

Lindemann et al. (2023) and Huaman et al. (2022) model datasets in Ontolex-Lemon (McCrae et al., 2017) on Wikibase. We will mention a few use cases dealing with Quechua, Latin, Kurdish, Basque and Serbian lexical and corpus data. These experiments aim at developing best practices for similar

projects in the future. The deployed models and workflows are entirely based on free software; this makes conversion and publishing of lexicographical data as LOD accessible to linguistic minorities on the Web. Furthermore, Wikibase instances can be straightforwardly federated with Wikidata, that is, Wikibase content, including lexeme descriptions, can be either transferred to Wikidata, or combined with Wikidata content in federated SPARQL queries, as long as Wikibase entities are provided with the identifier of the equivalent Wikidata entity. Linking lexemes to Wikidata, or transferring lexeme descriptions to Wikidata integrates own lexical data into the multilingual lexemes collection on Wikidata (cf. Nielsen 2020), which on the level of dictionary senses is itself linked to ontological concepts on the same platform.

The approach tested in this task is, in fact, a chance for under-represented communities to promote their language on their own Wikibase. Starting to curate digitised lexicographical datasets, which often contain inconsistent or noisy data, using an own Wikibase instance as an editing platform before going to Wikidata has several advantages. One advantage is to avoid creating uncertain or noisy material on Wikidata, and expecting an undefined community to curate it. After a community specialised in their own language's data is trained on their own Wikibase, members would most probably go on enriching their language's description on Wikidata, when the data is transferred to that global platform.

Stanković et al. (2023) generated Linked Data text corpora for 10 languages from ELTeC (European Literary Text Collection) using the NLP Interchange Format (NIF). The annotated version of the 1000 novels (100 per language) from POS, lemma and NER annotated TEI format was transformed into NIF, an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources, and annotations.

Additionally, several lexicons for under-resourced languages were published using Ontolex-lemon. Among these: Sentiment lexicon, football dictionary, Lithuanian-English cybersecurity terminology. Aspect verb lexicon for closely related languages (Stanković et al. 2024), generative Maltese verbal lexicon (Ionov and Rosner 2022).

### Progress as of M54

a. **LODifying lexical data using Wikibase and Wikidata: Use cases**

*Qichwabase*

Qichwabase (https://qichwa.wikibase.cloud/wiki/Qichwabase), a project started at the 2022 summer datathon (see section 4.2 of this report), and awarded the best project prize at that event, contains descriptions of 22,866 lexemes; these stem from the Runasimi Dictionary (https://runasimi.de/runaengl.htm), a multilingual lexical resource accessible online as a downloadable CSV file, collecting lemmas in the orthographic standard of the language (Hanan Runasimi), and providing correspondences in six European and twenty-five different languages of the Quechua family.

The contents of the source CSV file were modelled as instances of the core classes of the OntoLex Lemon model, the standard deployed on Wikibase for the representation of lexical data. One goal is to design working packages that are going to be released for the interested community to refine the data quality. Specific guidelines will enable the annotators to make informed decisions about, for example, aligning multilingual translation equivalents to word senses in entries describing polysemous words, since exact interlanguage correspondences between multiple word senses are not made explicit in the CSV source file. Annotators will use the Qichwabase graphical interface for their edits.

*Kurdish Wikibase*

Lindemann et al. (2023) transformed four resources freely available under an open source licence for Kurdish varieties (https://github.com/sinaahmadi/KurdishLexicography) using a Wikibase instance accessible at https://kurdi.wikibase.cloud.

In the cited paper, the authors point out some differences in the modelling of lexical data, according to the Ontolex-Lemon model, on one hand, and according to Ontolex as it is implemented by default in a Wikibase instance.

Kurdi Wikibase now contains large datasets describing three Kurdish varieties, with sense translations in English and/or Farsi, and is ready for a dedicated community to contribute.

*Basque language data on Wikibase*

Ahotsak Wikibase (https://datuak.ahotsak.eus/) Wikibase instance is used for experiments linking Basque dialectal lemmata and forms, and their attestations in video transcriptions from ahotsak.eus, to Basque lexemes on Wikidata. To this end, Basque lexemes descriptions from Wikidata have been combined with standard Basque lexeme forms as attested in ETC corpus. The resulting datasets remains ready for the definition of alignments between dialectal and standard forms, so that dialectal forms may inherit their morphological features descriptions from the latter.

Another instance, https://monumenta.wikibase.cloud/, includes entities that describe corpus tokens from a subset of the Basque Historical Corpus. These are annotated with links to their containing paragraph in the source document, e.g. a manuscript transcription on the Wikisource platform. In addition, Wikidata entities are deployed for semantic annotation. The historical text token is linked to standard dictionary entries on the same Wikibase instance (on lexeme, sense, and/or form level), which stem from the General Basque Dictionary. Finally, the data model proposed for this Wikibase instance also includes philological annotations, i.e. notes added by scholars to previous editions of the historical text.

*Latin data on Wikidata*

In the cited paper, Lindemann et al. (2023) describe a contribution to Wikidata, which consists of (a) the creation of a property that links Wikidata lexemes to the LiLa Knowledge Base (see https://lila-erc.org), and (b) the upload of 51,492 alignments between Wikidata and LiLa. For these lexemes, a federated query over the LiLa SPARQL endpoint (https://lila-erc.eu/sparql) already gives access to the wealth of information provided in the LiLa resources. Starting from a lexeme in Wikidata, for instance, it would be possible to retrieve all the occurrences of the words lemmatized under the connected lemma in the collection of LiLa corpora.

This contribution is an example of how own LOD datasets, in a Wikibase instance or other kind of LOD infrastructure, can be aligned to the Wikidata lexemes collection, which means a win-win situation, both for the Wikidata community, and for the provider of the external dataset, since the alignment ensures the findability and re-usability of the latter through the main Wikidata platform, i.e. through the graphical Wikidata interface as well as programmatically.

*ELTeC-NIF: European Literary Text Collection LLOD*

Stanković et al. (2023) generated Linked Data text corpora for 10 languages (German, English, French, Hungarian, Polish, Portuguese, Romanian, Slovenian, Spanish, and Serbian) from ELTeC (European Literary Text Collection) using the NLP Interchange Format (NIF). The annotated version of the 1000 novels (100 per language), in the so-called TEI level-2 format, was transformed into NIF, an RDF/OWL-based format that aims to achieve interoperability between NLP tools, language resources, and annotations. NIF provides support for part-of-speech tagging, lemmatization and entity annotation, enabling these three ELTeC level-2 layers transformation. Python code, including a

Jupyter notebook, is prepared for export from XML/TEI into NIF, available in colab subfolder in github repository. For Wikidata management mkwikidata library was used and for working with RDF rdflib. The implemented transformation pipeline is described, while the code and results ELTeC-NIF (https://llod.jerteh.rs/ELTEC/) are freely available for similar use cases. Serbian dataset is available on SPARQL endpoint (http://fuseki.jerteh.rs/#/dataset/SrpELTeC/query).

Metadata for selected novels are linked with already available in Wikidata, named WikiELTeC, as described in "From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)" by Ikonić-Nešić et al. (2022).

*Serbian Wikibase*

The Wikibase instance for Serbian (https://serbian.wikibase.cloud) is inspired by Qichwabase and other Wikibase instances; at the same time, this experiments shifts the use case of Wikibase from lexical to corpus data, aiming at representing both as part of the same dataset, including links from corpus tokens to dictionary entries, that is, on the Serbian Wikibase graph, to nodes representing lexemes, lexeme senses and lexeme forms. Nodes representing corpus tokens are annotated, e.g. with links to nodes representing named entities, which could be aligned to Wikidata entities. The proposed model is being populated with data from the Serbian literature corpus SrpELTeC () in NIF format (Stanković et al., 2023), explained in detail in the following, and with the Serbian SrpMD morphological lexicon in Ontolex-Lemon (Stanković et al. 2018).

In line with latest trends in the field of Linguistic Linked Open Data, a SrpELTeC token is modelled as a node in a LOD graph. Annotations to the token follow the NIF Ontology (Hellmann et al. 2013); this includes most prominently part of speech tag strings (associated to OLiA lexical categories), and lemma strings. The token string, the OLiA category, and the lemma string annotation are used for defining candidate dictionary entries for an alignment. A named entity layer available as part of SprELTeC includes seven classes (PERS, ORG, LOC, ROLE, EVENT, DEMO, WORK) that are mapped to OLiA, Wikidata and DBpedia.

The dictionary to link to, the SrpMD Serbian dictionary, is available in a format following the Ontolex-Lemon model (McCrae et al. 2017), which can be almost seamlessly transferred to a Wikibase (Lindemann et al. 2023).

The goal of this experiment, a novelty on Wikibase, is to obtain feedback on the proposed model, following requirements summarised as follows: to link corpus tokens (1) to a lexeme node, which is annotated with the standard lemma, and (2) to a lexical form associated to that lexeme, which can be annotated with the grammatical features describing the form. Additionally, the model caters for linking of corpus tokens to (3) a lexical sense associated to the lexeme, which can be annotated with a sense gloss, and (4), to an ontology concept the dictionary sense refers to.

*Verbal aspect database*

Stanković et al. (2024) created a database for verbal aspect for BSC, a group of closely related languages: Bosnian, Serbian and Croatian.

*Maltese generative lexicon*

Ionov and Rosner (2023) created a generative lexicon for Maltese verbs using OntoLex-Morph.

**Future plans**

- Within COST Action UniDive - WG2: “Lexicon-corpus interface” is aiming at interlinking MWE lexicon entries with their occurrences in corpora, and cross-lingually unified lexicography of idiosyncratic constructions will include linked data publication, as one of the proof-of-concept lexical encoding of MWEs and focus

on information integration. Serbian Wikibase (see above) is to be continued in this framework as an experiment for making use of the Wikibase software for the joint representation of corpus data and lexicographical data, including the linking of entities of both to each other.

- Within European Network On Lexical Innovation (ENEOLI) COST Action - low resourced language linked data will be included in research in WG1: Multilingual glossary of neology and WG2: Methodologies, digital resources and tools for neology. The multilingual glossary will be modelled, edited, and published using the Wikibase software.
- Publication of Ontolex Morph and FrAC (in collaboration with T1.1) with case studies in under-resourced languages.
- Finalisation and publication of W3C Best practices / guidelines LLOD for Under-resourced languages
- Collaboration with the DFG project “Open Text Collections” to produce RDF versions of glossed texts for a number of under-resourced languages using the Ligt vocabulary.

# 3. Guidelines and Best Practices on Linguistic Linked Open Data

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.” 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). More information about the Guidelines and best practices efforts of WG1 can be found in the respective deliverable available at: https://nexuslinguarum.eu/wp-content/uploads/2024/03/Deliverable-D1.4-Guidelines-and-Best-Practices-on-LLOD.pdf

# 4. Organised Events

Working Group 1 members have participated in the organisation of a number of events under the umbrella of the NexusLinguarum COST Action. Below we provide a brief summary of the organised events.

## 4.1 1st Workshop on PROfiling LINGuistic KNOWledgE gRaphs 2022

**Date:** May 23, 2022 , **Location:** Jerusalem, Israel

The focus of this workshop is to reveal novel approaches, methodologies and frameworks on profiling Linguistic Linked Data (LLD) (corpora, lexicons, ontologies, etc.) as well as to highlight tools and user interfaces that can effectively assist different use cases for profiling such data. In addition, the workshop seeks methodologies that help effective profiling in building real-world Linked Data applications leveraging linguistic data, as well as use cases that reveal success stories or aspects that have been neglected so far. The benefits of addressing Linguistic Linked Data profiling issues will not only help in understanding and exploring such data, but also provide the means to increase Linguistic Linked Data consumption, and to maintain track of the evolution of the relevant datasets.
More information at:
https://nexuslinguarum.eu/project/workshop-on-profiling-linguistic-knowledge-graphs-prolingknower/

## 4.2 Summer Datathon on Linguistic Linked Open Data 2022

**Date:** May 30 - June 3 2022**, Location:** Cercedilla, Spain

The 4th Summer Datathon on Linguistic Linked Open Data (SD-LLOD-22) was held physically from May 30th to June 3rd 2022 at Residencia Lucas Olazábal of Universidad Politécnica de Madrid, Cercedilla, Madrid (arrival expected on 29th evening). The main goal of the SD-LLOD-22 datathon is giving people from industry and academia practical knowledge in the field of Linked Data applied to Linguistics. The final aim was to allow participants to migrate their own (or other's) linguistic data and publish them as Linked Data on the Web and/or develop applications on top of Linguistic Linked Data. More information at:
https://nexuslinguarum.eu/project/4th-summer-datathon-on-linguistic-linked-open-data-sd-llod-22/

## 4.3 Summer Datathon on Linguistic Linked Open Data 2023

**Date:** June 11 - 16, 2023, **Location:** Lužnica, Croatia

The 5th Summer Datathon on Linguistic Linked Open Data (SD-LLOD-23) was held physically from June 11th to June 16, 2023 at Castle Lužnica, Croatia. It has the main goal of providing practical knowledge to people from industry and academia in the application of Linked Open Data technology to Linguistics and Language Technology. The ultimate goal is to enable participants to migrate their own (or other's) linguistic data and publish them as Linked Data on the Web and/or develop applications on top of Linguistic Linked Data (LLD). One of the main focus points this year will be the use of deep learning and neural approaches to/from LLD. More information at:
https://nexuslinguarum.eu/project/5th-summer-datathon-on-linguistic-linked-open-data-sd-llod-23/

## 4.4 2rd Workshop on PROfiling LINGuistic KNOWledgE gRaphs 2023

**Date:** September 12, 2023, **Location:** Vienna, Austria

In the last decades, we have experienced a substantial increase of Knowledge Graphs (KGs) published on the Web. The focus of this workshop is to reveal novel approaches, methodologies and frameworks on profiling Linguistic Linked Data (LLD) (corpora, lexicons, ontologies, etc.) as well as to highlight tools and user interfaces that can effectively assist different use cases for profiling such data. In addition, the workshop seeks methodologies that help effective profiling in building real-world Linked Data applications leveraging linguistic data, as well as use cases that reveal success stories or aspects that have been neglected so far. The benefits of addressing Linguistic Linked Data profiling issues will not only help in understanding and exploring such data, but also provide the means to increase Linguistic Linked Data consumption, and to maintain track of the evolution of the relevant datasets. More information at: https://prolingknower.disco.unimib.it/

## 4.5 4th Conference on Language, Data and Knowledge 2023

**Date:** September 12-15, 2023, **Location:** Vienna, Austria

Language, Data and Knowledge (LDK) aims at bringing together researchers from across disciplines concerned with the acquisition, curation and use of language data in the context of data science and knowledge-based applications. With the advent of the Web and digital technologies, an ever-increasing amount of language data is now available across application areas and industry sectors, including social media, digital archives, company records, etc. The efficient and meaningful exploitation of this data in scientific and commercial innovation is at the core of data science research, employing NLP and machine learning methods, as well as semantic technologies based on knowledge graphs. *Note: while the WG1 has not explicitly organised the conference, many WG1 members have participated in the organisation activities.* More information about the conference can be found at: http://2023.ldk-conf.org/

## 4.6 Wikibase-lex: LODifying lexical data using Wikibase 2023

**Date:** September 13, 2023, **Location:** Vienna, Austria

This half-day workshop was organised by David Lindemann, connected to WG1 task 1.5. This event was co-located to the 2023 LDK conference (see 4.5 above); Nexus Linguarum supported the event sponsoring one invited speaker's travel expenses, and providing the local infrastructure, as part of the workshop programme of the LDK conference.
At the workshop, Wikibase users reported on their projects to migrate lexicographical datasets to a Wikibase instance. This included the projects mentioned in section 2.5 above, regarding Quechua, Basque, Kurdish, and Latin data. Additional presentations were made by

Wikimedians active in the work on the Wikdiata lexems collections, about workflows and relevant software tools. The closing speech was held by Lydia Pintscher, head of the Wikidata team at Wikimedia Deutschland. All speeches were recorded; presentation slides and videos are available at the workshop website, https://wikibase-lex.sciencesconf.org/.

### 4.7 9th Workshop on Linked Data in Linguistics 2024

**Date:** May 25, 2024, **Location:** Torino, Italy

The Linked Data in Linguistics (LDL) workshop series has established itself as the premier venue for discussing the application of Semantic Web technologies to the fields of linguistics, digital lexicography, and digital humanities (DH).
While recent years have witnessed a steady growth in adoption of the technology in these areas, its uptake in other relevant domains, most notably in the case of natural language processing (NLP), continues to lag behind. This year, aside from embracing the full bandwidth of applications of LLOD technologies and the closely related area of knowledge graphs in linguistics, we welcome contributions addressing the application of LLOD technologies to NLP applications, as well as those dealing with emerging hot topics of future bridges between structured (linguistic) knowledge and neural methods. *Note that the workshop will happen a month after the Action ends, however, all the organisation and preparations of the workshop have happened during the lifetime of the NexusLinguarum Action.* More information at: https://nexuslinguarum.eu/project/9th-workshop-on-linked-data-in-linguistics-ldl-2024/

## 5. Future Directions and Summary

Within the 54 months of the NexusLinguarum COST Action, WG1 members developed a solid foundation for further collaborations. The work within WG1 has fulfilled its set goals, and in particular that of establishing firm foundations for the development, publishing, modelling, linking, enrichment, quality assurance and repair of LLOD resources. WG1 has also organised a number of events in order to strengthen the links among the community members. Also, several STSMs and Virtual Mobility grants have been successfully organised. After the Action, the WG1 members will definitely continue their collaboration within the i) Ontolex W3C community group by working on additional Ontolex modules, ii) BPMLOD W3C community group by contributing to the best practices efforts, iii) LD4LT community group on development of Linked Data methods for linguistic technologies, iiii) organising further editions of events which have already taken place (i.e. workshops, training schools, conferences). Last but not least, WG1 members will continue their collaboration within the scope of the UniDive and ENEOLI COST Actions.

# 6. WG1 publications


- di Buono, M. P., Spahiu, B., & Barbu Mititelu, V. (2024). Evaluating Large Language Models for Linguistic Linked Data Generation. In *Proceedings of the First Workshop DLnLD: Deep Learning and Linked Data-LREC-COLING 2024-The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*.

- 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.

- Barbu Mititelu, V., Di Buono, M. P., Oliveira, H. G., Spahiu, B., & Valūnaitė-Oleškevičienė, G. (2023, September). Adopting Linguistic Linked Data Principles: Insights on Users' Experience. In *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 347-357).

- Blerina Spahiu, Renzo Alva Principe, and Andrea Maurino.- Profiling Linguistic Knowledge Graphs.- *Proceedings of the 4th Conference on Language, Data and Knowledge*. 2023
- Gollam Rabby, Farhana Keya, Vojtech Svátek, Blerina Spahiu - Pruning and re-ranking the frequent patterns in knowledge graph profiling using machine learning.- *Proceedings of the 4th Conference on Language, Data and Knowledge. 2023.*
- Michael Rosner, Sina Ahmadi, Elena-Simona Apostol, Julia Bosque-Gil, Christian Chiarcos, Milan Dojchinovski, Katerina Gkirtzou, Jorge Gracia, Dagmar Gromann, Chaya Liebeskind, Giedrė Valūnaitė Oleškevičienė, Gilles Sérasset and Ciprian-Octavian Truică, Cross-lingual link discovery for under-resourced languages. In Proceedings of the Language Resources and Evaluation Conference, pages 181–192, Marseille, France, June 2022. European Language Resources Association.
- Maxim Ionov and Mike Rosner. 2023. Beyond Concatenative Morphology: Applying OntoLex-Morph to Maltese. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 385–391, Vienna, Austria. NOVA CLUNL, Portugal.
- Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos1, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação, Silvano, Blerina Spahiu, Andrius Utka, Ciprian-Octavian Truica, Giedrė Valūnaitė Oleškevičienė, Multilinguality and LLOD: A Survey Across Linguistic Description, Semantic Web Journal (to appear 2024)

- Dimitar Trajanov, Elena-Simona Apostol, Radovan Garabík, Katerina Gkirtzou,Dagmar Gromann, Chaya Liebeskind, Cosimo Palma, Michael Rosner, Alexia Sampri, Gilles Sérasset, Blerina Spahiu, Ciprian-Octavian Truică, Giedre Valunaite Oleskeviciene, From Linguistic Linked Data to Big Data, Proceedings of LREC 2024, Torino, 2024

- Michael Rosner and Maxim Ionov, Linguistic LOD for Interoperable Morphological Description, Proc 9th Workshop on Linked Data in Linguistics, LREC 2024, Torino, May 2024

- Anas Fahad Khan, Maxim Ionov, Christian Chiarcos, Laurent Romary, Gilles Serasset, Besim Kabashi. On Modelling Corpus Citations in Computational Lexical Resources. In Proceedings of LREC 2024, Torino, 2024

- Florentina Armaselu, Elena-Simona Apostol, Christian Chiarcos, Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, Andrius Utka, Giedrė Valūnaitė-Oleškevičienė, Towards a Conversational Web? A Benchmark for Analysing Semantic Change with Conversational Knowledge Bots and Linked Open Data. In Proceedings of the 4th Conference on Language, Data and Knowledge, 2023.

- Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros Muñoz, Ciprian-Octavian Truică. When linguistics meets web technologies. Recent advances in modelling linguistic linked data. In Semantic Web, vol. 13, no. 6, pp. 987-1050, 2022.

- Fahad Khan, Christian Chiarcos, Thierry Declerck, Maria Pia Di Buono, Milan Dojchinovski, Jorge Gracia, Giedrė Valūnaitė-Oleškevičienė, Daniela Gifu. 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, 2022.

- Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. "Sentiment Analysis of Serbian Old Novels." In Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data, pp. 31-38. 2022.

- Nešić-Ikonić, Milica, Ranka Stanković, Christof Schöch, and Mihailo Škorić. "From ELTeC Text Collection Metadata and Named Entities to Linked-data (and Back)." In Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference, pp. 7-16. 2022.

- Chiarcos, Christian, Ranka Stanković, Maxim Ionov, and Gilles Serasset. "Bridging Computational Lexicography and Corpus Linguistics: A Query Extension for OntoLex-FrAC." In LREC-COLING 2024. 2024.

- Stanković, Ranka, Christian Chiarcos, Miloš Utvić, and Olivera Kitanović. "Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data." In Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 180-191. 2023.

- Lazarević, Jelena, Ranka Stanković, Mihailo Škorić, and Biljana Rujević. "Football terminology: compilation and transformation into OntoLex-Lemon resource." In Proceedings of the 4th Conference on Language, Data and Knowledge, pp. 634-645. 2023.