D1.3

Intermediate Activity Report

Working Group 1 “Linked data-based language resources”

Main authors:

Milan Dojchinovski and Julia Bosque-Gil

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science
COST Action	18209
Starting Date	26 October 2019
Duration	48 months
Project Website	https://nexuslinguarum.eu/
Chair	Jorge Gracia
Main authors	Milan Dojchinovski, Julia Bosque-Gil
Contributors	Sina Ahmadi, Verginica Barbu Mititelu, Christian Chiarcos, Maria Pia di Buono, Maxim Ionov, Anas Fahad Khan, Hugo Gonçalo Oliveira, Blerina Spahiu, Vojtěch Svátek, Mike Rosner
Reviewer	NexusLinguarum core group team
Version | Status	final
Date	31/10/2021

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 first 24 months of activity. Over the course of these two years WG1 has successfully managed to establish a stable leading structure and both inter-WG and intra-WG collaborative dynamics. The WG has fulfilled its defined goals for the first half of the Action, namely: starting to provide 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 organization of events such as training schools and workshops, the execution of Short Time Scientific Missions and Virtual Mobility Grants, and the dissemination of the resulting work via the organization of a special issue.

1. Introduction

Language resources (LRs) play a key role in research in humanities and in the development of NLP applications. Working Group 1 (WG1) aims to bring the benefits of the linked (open) data (LOD) paradigm to these areas in order to make them more easily discoverable, reusable and interoperable both with one another and with the tools consuming them. Specifically, WG1 is concerned with layering the foundations and developing best practices for the evolution, creation, improvement diagnosis, repair and enrichment of Linguistic LOD resources and value chains.

The first few months of the Action were dedicated to structuring and planning the work within WG1. From an organizational perspective, first, working group (co-)leaders have been elected. Next, an open call for task (co-)leader nominations has been announced during which task (co-)leaders have been elected. Immediately after the 1st plenary meeting in January 2020 in Prague, Czech Republic, the WG1 members have agreed on a date/time for regular coordination calls which are currently running each month on Friday at 10 a.m. CEST in the third week of the month. For the first half period of the NexusLinguarum CA the WG1 has organised over 20 regular calls. As of October 2021, WG1 consists of over 78 members.

The work within WG1 is organised into five tasks, each dedicated to a particular aspect of the Linguistic Linked Data lifecycle. Task 1.1 focused on LLOD modelling (see Section 2.1), Task 1.2 focused on creation and evolution of LLOD resources in a distributed and collaborative setting (see Section 2.2), Task 1.3 focused on cross-lingual data interlinking, access and retrieval in the LLOD (see Section 2.3), Task 1.4 aiming at improving and monitoring quality of LLOD sources (see Section 2.4) and Task 1.5 dedicated to development of the LLOD cloud for under-resourced languages and domains (see Section 2.5). Each task is led by a task leader and a task co-leader, with expertise profiles complementing each other. Table 1 summarises this structure.

While the WG1 tasks advance in parallel, joining efforts periodically depending on the goals defined in roadmaps for each grant period and agreed within WG1, Task 1.5 is particularly suitable for collaborative work with other tasks. The reason for this lies in the fact that features related to under-resourced languages might affect each of the stages of the LRs generation cycle, from modelling to the quality assessment of the resulting resource.

During these past two years, there were a number of events supported by NexusLinguarum and with particular relevance to WG1.

The first online training school in the context of the Action, Introduction to Linked Data for Linguistics, was an online event taking place on February 8-12, 2021. The training school aimed at promoting and teaching the foundations of linguistic data science and its related technologies to people from both academia and industry, and was organised under the umbrella of the EUROLAN series of Summer Schools. See Section 4.1 for more information on the training school.

Role	Person	Country
WG1 leader	Milan Dojchinovski	Czech Republic
WG1 co-leader	Julia Bosque-Gil	Spain
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 and co-leader	Mike Rosner	Malta
	Sina Ahmadi	Ireland
Task 1.4 leader and co-leader	Blerina Spahiu	Italy
	Vojtech Svatek	Czech Republic
Task 1.5 leader and co-leader	Hugo Gonçalo Oliveira	Portugal
	Max Ionov	Germany

Table 1. Structure of WG1 (as of October 31st, 2021).

The WG1 members also actively participated in the organization of the *Language, Data and Knowledge Conference (LDK) 2021*, which took place on September 2nd and 3rd in Zaragoza, Spain. WG1 members have contributed with papers presented at the main conference but also WG1 has been involved in the organization of several WG1 related events co-located with LDK: *the 3rd W3C Ontolex Workshop* (see Section 4.2) and *the W3C LD4LT Annotation Workshop* (see Section 4.3), which are well aligned with the Task 1.1 activities.

WG1 members have also organised a special issue in the Semantic Web Journal on "Latest Advancements in Linguistic Linked Data". The special issue aims at gathering high-quality contributions, supported by a robust evaluation, which present an advancement in the state-of-the-art in the field of LLD methodologies and technologies and their use for NLP and provide insights into the new challenges ahead. For more information about the special issue see Section 5.1.

WG1, and in particular Task 1.2 and Task 1.5 co-leaders supported with few other NexusLinguarum members, have published a policy brief on the inclusion of data from under-resourced languages. More information about the policy brief can be found in Section 2.5.

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

- D1.1 Report and Training Materials of the 1st Training School
- D1.2 Policy brief about the inclusion of data from under-resourced languages
- D1.3 Intermediate and final activity report (this document)

STSMs and VM grants have also provided the opportunity to boost the collaboration among WG1 members and directly contribute to the goals and the developments of particular WG1 tasks. More specifically, three WG1 related STSMs and one Virtual Mobility grant have been

executed. More information on the STSMs and the Virtual Mobility grants can be found in Section 3.

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 gives an overview of the STSMs and Virtual Mobility grants which have been executed and are related to the WG1 goals. Section 4 provides a summary of the organised WG1 related events. Section 5 provides information about WG1 related activities such as the organised special issue in the Semantic Web Journal and the interactions with the other WGs. Section 6 outlines the future directions of the WG1 and provides an overall summary of the report. Finally, Section 7 lists the outcome publications generated by WG1 members in the context of the different tasks.

2. Tasks Reports

2.1. Task 1.1 LLOD modelling

Task Leaders:

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

General Overview

The overall goal of this task is to investigate how to model linguistic resources/phenomena/tools using Semantic Web standards, such as the Resource Description Framework (RDF), in order to facilitate the linking, integration and (re)usability of language resources modelled and published as Linguistic Linked (Open) Data (LL(O)D). More precisely, this task seeks to provide answers to the following questions:

- How to use RDF to make linguistic data and tools for working with that data more accessible, more interoperable and usable?
- Which RDF vocabularies exist for language resources, how are they currently being used and what needs improvement in terms of e.g., coverage?
- Whether sufficient resources currently exist for bridging RDF and non-RDF technology/standards (TEI[1], ISO-LMF[2], etc.) for language resources and if not how do we define them?
- What are the benefits of RDF as a data framework and how can disadvantages be compensated?

These questions are primarily discussed in a number of parallel telcos with narrower focus on specific types of resources (esp. lexical resources and linguistic annotations), and requirements (esp. within lexical resources).

Progress as of M24

[1] https://tei-c.org/
[2] https://www.iso.org/standard/68516.html

- State-of-the-Art analysis: An extensive survey was carried out within Task 1.1 covering use of specific LL(O)D vocabularies for different categories of language resources and what gaps still exist, as well as the use of these vocabularies within various prominent projects and initiatives. This survey resulted in an approx 60 page article[3] which was submitted to a special issue of the Semantic Web Journal on Linguistic Linked Data and is still under review.
- Community work: In this regard, Task 1.1 has focused on the facilitation of information flow (especially in terms of requirements) between NexusLinguarum participants, tasks and use cases and ongoing discussions within W3C Community Groups (CGs) OntoLex[4] and LD4LT.[5] Both task leads are co-chairs and active participants in both communities. Contact has also been made with specialists working on the Lexical Markup Framework (LMF)[6] and the Text Encoding Initiative (TEI)[7] in order to work on potential crosswalks between OntoLex and these standards. In addition, the co-leader of Task 1.1 and two other participants of the task (Penny Labropoulou and Marco Passarotti) helped to organise a CLARIN event ("CLARIN Café on Linguistic Linked Data") dedicated to initiating a closer dialogue between the linguistic linked data community and CLARIN ERIC. Speakers at the event included the Task 1.1 leader and other participants of the COST action[8].
- Vocabulary development: The development of vocabularies has concerned three major lines of work, namely:
    - Vocabulary development for lexical resources (dictionaries, lexical networks, term bases) in collaboration with W3C Community Group Ontology-Lexica (OntoLex). In particular, the following examples of collaboration can be singled out:
        - Joint telcos (multimodality: together with the OntoLex group and NexusLinguarum Task 3.4 (see Section 5.2.1); OntoLex-FrAC[9]: frequency, attestation and corpus-based information in lexical resources; OntoLex-Morphology[10]: co-led with NexusLinguarum Task 1.5)
        - Joint workshop (including an OntoLex workshop at LDK-2021 (see Section 4.2) and publications.
    - Vocabulary development for linguistic annotation (NLP web services, corpus technology) in collaboration with W3C Community Group Linked Data for Language Technology (LD4LT). This collaboration has been centred around the following points:
        - Joint telcos, so far focusing on requirement analysis
        - Survey on requirements for vocabularies for linguistic annotations on the web

[3] http://www.semantic-web-journal.net/content/when-linguistics-meets-web-technologies-recent-advances-modelling-linguistic-linked-open-0
[4] https://www.w3.org/community/ontolex/
[5] https://www.w3.org/community/ld4lt/
[6] http://www.lexicalmarkupframework.org
[7] https://tei-c.org
[8] https://www.clarin.eu/event/2021/clarin-cafe-linguistic-linked-data
[9] https://www.w3.org/community/ontolex/wiki/Frequency,_Attestation_and_Corpus_Information
[10] https://www.w3.org/community/ontolex/wiki/Morphology

        - Joint workshop on "Harmonizing Linguistic Annotations" at LDK-2021 (see Section 4.3). This focused on providing (and recording) background descriptions as well as use case descriptions.
    - Proposals for new OntoLex modules and ideas for additional potential modules. These proposals and suggestions have arisen from collaboration between NexusLinguarum tasks and OntoLex, including a planned terminology module as well as specific requirements for modelling discourse phenomena (UC 4.2.2) and spatio-temporal data (UC 4.2.1), as well as within LD4LT, including a suggestion on Fragment Identifiers.

Future plans

Future plans include:

- Continuing collaboration between NexusLinguarum and the W3C Community Groups with the intention of bringing to publication of OntoLex-FrAC and OntoLex-Morphology.
- Further discussions as to how to bridge between LLOD standards and other, pre-existing standards, including:
    - Potential further development and testing of a XSLT transformation between TEI-LEX 0[11] and OntoLex-Lemon[12] on the basis of an already existing transformation devised by John McCrae and Laurent Romary;
    - Continued discussions with the drafters of the ongoing multipart version of LMF with the ISO working group, ISO/TC 37/SC 4/WG 4, to establish a closer alignment and harmonization between OntoLex and LMF
- Further discussions on how to create more synergies and collaborations between the LLD community and research infrastructures such as CLARIN especially with regard to LLD models and vocabularies
- Follow on work from the SOTA survey to see how the gaps which were identified (in terms of the need for new models for different kinds of language resources) can be filled in.

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. New approaches will be analysed to the distributed and collaborative creation and extension of LRs that allow parties to easily extend existing resources and publish their own extensions as LD.

Progress as of M24

[11] https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html
[12] https://www.w3.org/2016/05/ontolex/

- State-of-the-Art of linked data (LD) resources: 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[13] and Annohub[14]) 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"[15], submitted to 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[16].
- Survey on the obstacles for LD reuse, extension, and creation[17]. 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) [18], we have made the decision to rethink the strategy of attracting more participants.

Synergies with other tasks: during all these two years, 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

For the near future:

- Reach out to more language resources developers and users to have a clearer understanding of the obstacles in the creation, publication, reuse, linking of resources;

[13] https://lod-cloud.net
[14] https://annohub.linguistik.de/
[15] http://www.semantic-web-journal.net/content/paving-way-enriched-metadata-linguistic-linked-data-0
[16] https://github.com/unior-nlp-research-group/melld
[17] https://forms.gle/aaCHV1fsxM9CbJjA7
[18] https://docs.google.com/spreadsheets/d/102XbYrcVnw-A4bo-D7MhorPX4WCKnohHYSHKZlHMwR8/

- Publish the results of the survey.

Long-term goals:

- Focus on the evolution of language resources;
- Propose a methodology to ease the (re)use and the creation of LLD in a collaborative setting, which takes into account the integration of available tools to formalize LRs.

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

Task Leaders:

- Leader: Mike Rosner
- Co-leader: Sina Ahmadi

General Overview

Within WG1, the task 1.3 focuses on cross-lingual data interlinking, access and retrieval in the LLOD by identifying and studying novel (semi-) automatic methods that help increasing the interlinking across LLOD datasets, as well as methods and techniques based on LLOD for accessing and exploiting data across different languages. To do so, we provide a general overview of *cross-linguality* versus *multilinguality*, two concepts which are related and of importance to LLOD, as follows:

- Multilinguality is a feature of a resource based on the presence of content in at least two languages. In other words, a multilingual resource is characterized as one containing information in more than one language. Examples of multilingual resources include:
    - Multilingual BERT - Pretrained multilingual language model based on the top 104 languages with the largest Wikipedia using a masked language modelling (MLM) objective[19]. See also RoBERTa and XLM-R.
    - A website which includes versions of pages in more than one language allowing the user to select the language of presentation
    - A codeswitching corpus whose sentences include segments in different languages
    - A multilingual terminology containing equivalent terms in more than one language e.g. IATE, the terminology database of the EU.
- Cross-linguality is typically applicable to tasks or operations in which items expressed in one language come to be associated with items expressed in another language. For example
    - Cross-lingual information retrieval (CLIR), i.e. retrieval of relevant documents that are written in a language other than the language in which the query is expressed.

[19] Strictly BERT encompasses an approach, a tool as well as data so it is not merely a data resource

- Word translation: retrieval of a word in a target language that is semantically equivalent to a given word in a source language.
- Creation of explicit links between source and target words to support the task of word translation.

Therefore, a cross-lingual resource is characterized by the presence of links or equivalences between data in different languages that allow navigation from information in one language to information in another language. These cross-lingual links could have been established at the time of creation of the resource or at a later stage (manually or automatically discovered). A resource, an environment or an approach that allows this traversal from one language to another can thus be considered cross-lingual.

Progress as of M24

- Development of a framework for analyzing cross-linguality: Once the key features of cross-linguality were defined, we focused on resources for which cross-lingual links are manually or automatically created. To this end, we have been developing a framework for breaking down the main theme into the following four more manageable subparts as follows:
  a. Cross-lingual Link Discovery
  b. Cross-lingual Link Representation
  c. Cross-lingual Link Storage and Reuse
  d. Cross-lingual Access, Retrieval and Extraction
- Literature review: Accordingly, the WG1.3 subgroup has been scanning the literature (partly in collaboration with WG 3.3) creating summaries of what has been achieved with respect to these subtasks.

Future plans

The focus of our present work is to publish a position paper whose main aims are to

- Identify the key challenges that affect each of the above sub-themes , and
- suggest some promising future directions which address the most important of these challenges.

Ideally, some of the directions suggested could be pursued within subsequent phases of NexusLinguarum.

2.4. Task 1.4 Improving and monitoring quality of LLOD sources

Task Leaders:

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

General Overview

The overall goal of this task is to evaluate the quality of linguistics resources in the LOD cloud. More specifically, this task aims to provide a general overview of the quality of the LLOD resources, based on dimensions and metrics for LOD resources, and propose new metrics specifically for linguistic ones. Moreover, as a pre-step for data quality assessment,

this task aims to provide an analysis of the content of LLOD by examining their content with data profiling techniques.

Progress as of M24

- State of the art of data quality: A thorough analysis was carried out in order to identify data quality dimensions and metrics that are used to estimate the quality of RDF resources. The list of the dimensions and the respective metrics have been identified. Moreover, possible tools and approaches that implement the list of defined metrics have been also identified.
- State of the art of data profiling: An extensive analysis of the approaches and tools that profile linguistics datasets has been carried out. In such analysis, we considered tools and approaches that not only provide information about the content of the dataset in the form of statistics but also those that extract patterns and summarise their content.
- Preprocessing of LLOD resources: All datasets belonging to the linguistics domain have been preprocessed. The datasets and their ontologies in the LLOD belong to different formats. Because (1) not all tools for analysis support any kind of format, and (2) datasets and their ontologies have different syntactic errors, all datasets have been preprocessed and converted into the N-Triples format, and to OWL for ontologies.
- Data profiling analysis: All datasets in the LLOD have been processed with the ABSTA[20] profiling tool (Alva Principe et al., 2021). For each, a profile has been extracted. The profile consists of (1) a summary with patterns about the relations in the data and (2) statistics about different features (classes, properties, data types, vocabularies, external classes/properties, etc.). Currently, we are analysing such profiles with the aim of providing a general overview of the characteristics of the datasets within the linguistics domain.
- Data quality analysis: Luzzu (Debbatista et al., 2016) is considered as the main tool for quality assessment. However, there are some bugs that the Luzzu maintainers are trying to resolve. Meanwhile, we are assessing those metrics for which no implementation problems have been identified. Moreover, the ABSTAT profiling tool also gives insights about some quality issues such as cardinality constraints (Spahiu et al., 2018). Currently, we are also testing other tools that might be a good fit for the quality assessment of LLOD data.

Future plans

Short-term plan:

- Publish the result of the analysis of the LLOD profiling phase: We intend to publish the results from the analysis of the content of the LLOD datasets in order to provide users and the linguistics community an overview of the characteristics of LLOD.
- Continue collaborating with Task 1.2 and Task 1.5 and other tasks and working groups within NexusLinguarum.

Long-term plan:

- Publish the survey about quality assessment of LLOD.
- Define and propose new metrics that are specific for linguistic data.

[20] "ABSTAT." http://abstat.disco.unimib.it/. Accessed 21 Oct. 2021.

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

Task Leaders:

- Leader: Hugo Gonçalo Oliveira
- Co-leader: Maxim Ionov

General Overview

The main goal of this task is to **evaluate**, **improve** and **promote** the presence of under-resourced languages and domains in the LLOD cloud. More specifically:

- To evaluate the current language distribution in the LLOD cloud;
- To chart possible ways for increasing coverage for languages and domains that are currently under-represented in the LLOD cloud;
- To promote using LD technologies among people working with under-resourced languages.

Given the "vertical" nature of the task, i.e. having intersections with all the other tasks in the WG1, most of the work is collaborative among members in different tasks.

Progress as of M24

- State-of-the-art of LD resources (in collaboration with Task 1.2 and Task 1.4): An overview of the status of languages and domains in the LOD cloud and Annohub, with a focus on the LLOD cloud, including an analysis according to linguistic resource types. Language representation in the LLOD cloud was further analysed, as well the types of linguistic resources, which resulted in insights on language representativeness and languages being under-resourced in the LLOD cloud, all described in the paper "Paving the Way for Enriched Metadata of Linguistic Linked Data", submitted to a special issue of the Semantic Web Journal on Linguistic Linked Data (under review at the time of submitting the report).
- Participation in designing the survey about LD knowledge and usage among language professionals (in collaboration with Task 1.2 and Task 1.4), described among Task 1.2 results.
- Vocabulary development for representing morphological data in RDF as an OntoLex module (in collaboration with Task 1.1). Participation in the discussions about morphological data representation to make sure that under-resourced languages are taken into account and the module will be applicable to morphological features that some of these languages possess.
- Vocabulary development and dataset conversion for language documentation data (in collaboration with Task 1.1). Publishing a new LD resource using an RDF-native vocabulary Ligt[21] as a showcase of including language documentation data in the landscape of LLOD resources.

[21] https://github.com/acoli-repo/ligt

- Policy brief on increasing the digital presence of under-resource languages in and for technological development.[22] Language technology development hinges greatly on the existence of language resources, including lexica, corpora, databases, etc. However, there are dramatic differences in the availability of such resources across languages. This report summarises some of the main technological, cultural and socio-economic barriers for a truly multilingual Europe and proposes a set of actions to take by different stakeholder profiles, namely: researchers and developers, language data providers, policy makers and funding agencies.

Future plans

- Further promotion of publishing and using LLOD resources within language professionals dealing with under-resourced languages: organising tutorials, summer schools and workshops with a focus on converting under-resourced language data;
- Further development of OntoLex modules in order to make sure under-resourced languages and phenomena occurring in them are considered while designing the vocabularies.
- Establishing cross-WG collaborations, including WG4 to address the topic of under-represented domains.

3. Short Time Scientific Missions and Virtual Mobility Grants

Until October 2021 several STSMs and one Virtual Mobility (VM) grant have been executed. The initial plan was to conduct more STSMs, however, due to the pandemic and the travel restrictions, the execution of STSMs has been very limited or even impossible. Besides all the limitations, the NexusLinguarum CA has managed to successfully support several grants, three of which are directly related to the goals of WG1. Below, we provide a brief summary of these WG1-related grants.

STSM-1: *Rights Management of Linguistic Linked Data (2021-07-12 to 2021-07-31).* The main goal of this STSM was to design representation methods of legal rights in the context of Linguistic Linked Data, with a focus on licenses for Language Resources and Technologies (LRTs) and licensing compatibility issues. Outcomes of this STSM, falling in the scope of Task 1.1, include: a draft ODRL profile for LRTs, a dataset of licenses with ODRL representations and a service for the transformation of licenses expressed in the Meta-Share ontology into ODRL policies.

STSM-2: *Generation, representation and exploitation of terminologies in the Semantic Web (2021-07-01 to 2021-09-30).* The mission consisted mainly in discussions to agree on and implement a first version of a standard model to represent terminologies following Semantic Web formats, specifically the OntoLex-Lemon model. The main result is a data model, represented in OWL, generated and validated with the ontology editor Protégé. The main

[22] https://nexuslinguarum.eu/results/policy-brief.

contributions of this STSM are in line with the objectives of WG1 specifically with task 1.1 on LLOD modelling.

STSM-3: *Corpus Analysis of Covid and health-related metaphors (2021-08-15 to 2021-08-30).* The aim of the STSM was the extension of multilingual resources, i.e. carrying out research on COVID-19 and health-related metaphors based on a multilingual corpus ParlaMint and creation of a corpus covering the COVID-19 pandemic situation as presented in the Lithuanian news and social media. The analysed and processed data was represented as an initial hierarchy/ontology of the frames obtained together with the related entries. The STSM contributes directly to Task 1.2.

In addition to the STSMs listed above, the NexusLinguarum CA executed a Virtual Mobility grant which has been directly related to the WG1 goals.

VMG-1: *Dictionary Metadata (06/09/2021 to 15/10/2021)*. The aim of this VM was to prepare the ground for the integration of metadata for printed and electronic dictionaries in LexBib, a digital bibliography and Knowledge Graph project for the domain of Lexicography and Dictionary Research, which currently stores metadata for lexicography-related publications. The VM resulted in the definition of a Dictionary Metadata (DM) model, analysing, re-using and extending the relevant vocabularies (META-Share ontology for the description of Language Resources, LexVoc vocabulary for lexicographic terms, and the FRBR model and BIBO ontology for bibliographic citations). This VM contributes directly to the objectives of Task 1.1 in vocabulary development.

4. Organised Events

4.1 The First Training school

The training school was held on February 8-12, 2021 and was aimed at students, academics, and practitioners to learn the foundations of Linguistic Data Science. During the course of the training school, the participants were introduced to a wide range of topics: from Semantic Web, RDF and ontologies, to modelling and querying linguistic data with state-of-the-art ontology models and tools. The training school has been organised under the umbrella of the EUROLAN series of Summer Schools and was hosted virtually (online) by two institutes of the Romanian Academy, the Research Institute for Artificial Intelligence in Bucharest and the Institute of Computer Science in Iași, as well as the "Alexandru Ioan Cuza" University of Iași, Romania. The training school was attended by 82 participants.

Due to the COVID-19 pandemic and current travel restrictions in Europe and beyond, the training school was held online. The training school has been driven and organised by WG1 members of the NexusLinguarum CA. Twelve lecturers were involved in the organization.

The training school provided valuable knowledge and trained a large number of computer scientists and linguists on how to work and benefit from linguistic linked data. This was the

first training school organised by the NexusLinguarum CA from the series of training events that are planned to take place. It aimed to serve as an introduction to the topic of linguistic data science and build the basis for the audience to attend subsequent training schools on more advanced topics during the Action's lifetime. All the materials created during the training school are publicly available and can be further used and utilized by the community. A detailed summary of the training school can be found in the dedicated report: D1.1. Report and Training Materials of the 1st Training School.

4.2 The 3rd Ontolex Workshop @ LDK 2021

The 3th Ontolex workshop[23], organised by the W3C Ontology Lexica community group, took place during the post-conference W3C day at LDK 2021 in Zaragoza (Spain) on 4th September 2021, in hybrid mode. The event consisted of a number of status update presentations on different topics related to Ontolex, and the different *lemon* modules, followed by open discussions. The goal was to get a good picture of the current state of Ontolex and to agree on future directions for both the model and the research community behind it. A number of issues regarding the representation and encoding of lexical resources as linked data were discussed, for instance in capturing and describing morphological aspects of lexical resources, as well as the latest advancements in encoding corpus information, frequency and attestations (FrAC). The future directions were also analysed, along with topics that the Ontolex group will focus on in the next years, with emphasis on the modelling needs perceived by the Ontolex community and which could be addressed in potential future modules, e. g. support for etymological description or multimodality.

4.3 LD4LT Annotation Workshop @ LDK 2021

At the W3C Day of Third Conference on Language, Data and Knowledge (LDK-2021), the W3C CG Linked Data for Language Technology (LD4LT) and NexusLinguarum Task 1.1 leads organised a joint workshop on *"Harmonizing Linguistic Annotations"*[24] with the support of many NexusLinguarum and LD4LT contributors on Sep 4, 2021. The goal of the workshop was to lay the groundwork for future consolidation of existing vocabularies by giving introductions to the most relevant standards, as well as elaborating on a number of specific use cases and requirements and summarising the current progress on requirement analysis for linguistic annotations on the web that was conducted within LD4LT and NexusLinguarum since 2019. The presentations have been collected and will be made publicly available via a designated website. We expect that on this basis, discussions in the coming months can shift from requirement analysis to concrete modelling proposals.

5. Other WG1 Related Activities

[23] https://www.w3.org/community/ontolex/wiki/3rd_Ontolex_Workshop_@_LDK_2021
[24] https://www.w3.org/community/ld4lt/wiki/LD4LT_Annotaton_Workshop_Zaragoza_2021

5.1 Special Issue on Latest Advancements in Linguistic Linked Data

With the rapid growth of the LLOD cloud and the increasing interest in the use of linked data for NLP, new challenges emerge concerning particular use cases and domain applications, language-specific features and quality dimensions, the evolution of LLD resources throughout time and the leverage of linguistic resources along LD technologies in NLP research, among other diverse aspects.

WG1 and WG2 members have thus organised a *special issue* in the Semantic Web Journal on the "Latest Advancements in Linguistic Linked Data" (call for papers). The guest editorial board invited high-quality contributions presenting advancements in the state-of-the-art in the field of LLD methodologies and technologies and their use for NLP. The list of topics are particularly relevant to the work carried out in the Action and included Knowledge Representation for Linguistic Data, LLD Generation and Evolution, LLD Publication, Querying and Visualization, LLD and NLP research, Applications and Use Cases.

The initial deadline for submission was the 20th of November 2020, but this was extended until January 25th, 2021. The special issue received a total of 13 submissions, 5 of which were directly related to NexusLinguarum efforts. The special issue is currently on the second round of peer-review.

5.2 Interaction with the other Working Groups

WG1 collaborates closely with other WGs in the CA. There are three main open lines of joint work with WG3, *Support for linguistic data science*, and WG4, *Use cases and applications*.

5.2.1 WG1 - WG3 Collaboration

WG3 is concerned with the applications of data analytic techniques at a large scale, together with LLOD and NLP techniques, to foster the study of linguistic data. Such support for linguistic data science needs to consider the nature of the different resources, too. Since WG1 is focused on the different stages of the LLOD life-cycle and the generation of LLOD-based LRs, the two WGs have joined forces to address challenging topics such as multimodality in linguistic data and cross-lingual interlinking.

Multimodality. In the context of Task 3.4 *Multidimensional linguistic data*, multimodality was identified as a case of multidimensionality. As a first use case, members of Task 1.1 and Task 3.4 are working on the description of sign language data categories and data as RDF, according to available vocabularies and on-going extensions. The analysis of other resources including pictographs and data not represented in Unicode is also foreseen for the next half of the Action.

Multilingual and cross-lingual linking. In the context of Task 3.3, *Linking structured multilingual language data across linguistic description levels,* Task 1.1, Task 1.2, Task 1.3 and Task 1.4 members are working together with WG3 in a systematic survey on the approaches and research on multilingual LLOD. In parallel, Task 1.3 is leading a position statement on the challenges for cross-lingual linking of linguistic resources, with a focus on link discovery, link representation, access, retrieval and storage.

5.2.2 WG1 - WG4 Collaboration

WG4 studies possible use cases and applications of the technologies involved in NexusLinguarum. Specifically, WG1 and WG4 have identified different aspects in which the requirements detected by WG4 for a specific use case can be addressed jointly with WG1. The work in WG3 for some of these use cases is highly relevant as well, which leads to lines of work involving these three WGs.

The most active line of collaboration across the two WGs as of today revolves around the topic of **diachronicity**. This collaboration concerns a specific WG4 use case, namely, UC4.2.1, the main goal of which is the creation of a comparative methodological framework for tracing the evolution of concepts across different languages and humanities disciplines (including history, literature and philosophy). This methodology foresees the construction of multilingual Semantic Web ontologies on the basis of results emerging from the application of NLP tools on diachronic corpora. These ontologies will trace changes in the significance of concepts and the linguistic expressions used to refer to them, which will likely necessitate the extension of currently existing LLD vocabularies in order to deal with diachronic information. The role of WG1 is to work on this and other modelling aspects of the use case. One initial result of this ongoing work has been the submission of a joint article for the special issue of the Semantic Web Journal on Linguistic Linked Data. This article is currently under review[25].

With respect to the next topics to explore jointly in the following months, a **matrix of collaboration** was put forward by WG4 and completed jointly with WG1. This matrix is summarised due to space reasons in Table 2.

[25]http://www.semantic-web-journal.net/content/llod-and-nlp-perspectives-semantic-change-humanities-research-0

	Task 4.1 Linguistics		Task 4.2 Humanities & Social Sciences		Task 4.3 Technology		Task 4.4 Life Sciences	
	UC4.1.1 Media and Social Media	UC 4.1.2 Language Acquisition	UC 4.2.1 Humanities	UC 4.2.2 Social Sciences	UC 4.3.1 Cybersecurity	UC 4.3.2 FinTech	UC 4.4.1 Public Health	UC 4.4.2 Pharmacy
T1.1 Modelling	Cross-linguistic modeling of hatespeech/offensive language taxonomies	LLOD modelling of language acquisition data	Modelling of semantic change in diachronic corpora	Modelling of discourse annotations and discourse marker inventories; semantics and pragmatics of speaker's attitudes	Modelling of cybersecurity terminology data		Modelling an ontology of conceptual metaphors through semantic frames	
T1.2 Resources	LRs related to a correlation between emotions/sentiment types and categories of offence						Ontologies related to public opinion or public sentiment	
T1.3 Interlinking	Interlinking of sentiment/emotion classes to offensive language categories		Interlinking of concepts accross different languages using multilingual diachronic corpora	Analysis of multilingual aspects of discourse markers; interlinking semantics of speaker's attitudes	Application of LLOD for bilingual termbase data linking			
T1.4 Sources quality								
T1.5 Under-resourced languages								

Table 2. Matrix of potential topics of collaboration for the next months designed by WG4 and filled in at an inter-WG meeting.

6. Future Directions and Summary

Within the first 24 months of the NexusLinguarum COST Action, the work within WG1 has been primarily focused on the establishment of a stable leading structure and kick-of and intensify both the inter-WG and intra-WG collaborations. The work within WG1 has fulfilled its set goals, and in particular: establish foundations for 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.

For the following period, the WG1 will put its focus on i) execution of more STSMs and Virtual Mobility Grants, ii) provide stronger support for ITC conference grants, iii) organise WG1 dedicated meeting on specific topics, iv) participate in the organization of planned NexusLinguarum events such as the training schools, v) intensify collaboration with other related initiatives in the field of standards and metadata, and last but not least v) provide well defined guidelines and best practices for development and use of Linguistic Linked (Open) Data resources.

7. Publications

WG1 (publications spanning across all tasks)

Ionov, M., McCrae, J.P., Chiarcos, C., Declerck, T., Bosque-Gil, J., Gracia, J., eds. Proceedings of 7th Workshop on Linked Data in Linguistics (LDL 2020). ELRA; 2020.

Kernerman, I., Krek, S., McCrae, J. P., Gracia, J., Ahmadi, S., & Kabashi, B. (2020, May). Proceedings of the 2020 Globalex Workshop on Linked Lexicography. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*.

Declerck T., McCrae J., Hartung M., Gracia, J., Chiarcos, Ch., Montiel, E., Cimiano, P., Revenko, A., Sauri, R., Lee, D., Racioppa, S., Nasir, J., Orlikowski, M., Lanau-Coronas, M., Fäth, C., Rico, M., Elahi, M. F., Khvalchik, M., Gonzalez, M., Cooney, K. Recent Developments for the Linguistic Linked Open Data Infrastructure. In: Proc. of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille (France): ELRA; 2020:5660-5667.

T1.1

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., García, E. G. B., Gracia, J., ... & Truica, C. O. (under review). When Linguistics Meets Web Technologies. Recent advances in Modelling Linguistic Linked Open Data. Semantic Web Journal.

[in collaboration with WG4] Armaselu, F., Apostol, E. S., Khan, A. F., Liebeskind, C., McGillivray, B., Truica, C. O., ... & van Erp, M. (under review). LL (O) D and NLP Perspectives on Semantic Change for Humanities Research. Semantic Web Journal.

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, F., Stolk, S., ... & McCrae, J. P. (2020, May). Modelling frequency and attestations for ontolex-lemon. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography (pp. 1-9).

Chiarcos, C., Declerck, T., & Ionov, M. (2021, January). Embeddings for the Lexicon: Modelling and Representation. In Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6) (pp. 13-19).

T1.2

[in collaboration with T1.4 and T1.5] di Buono, M. P., Oliveira, H. G., Mititelu, V. B., Spahiu, B., & Nolano, G. (under review) Paving the Way for Enriched Metadata of Linguistic Linked Data. Semantic Web Journal.

G. Nolano, M. Fazleh Elahi, M. Pia di Buono, B. Ell, P. Cimiano "An Italian Question Answering System based on grammars automatically generated from ontology lexica". CLiC-it 2021: Eighth Italian Conference on Computational Linguistics - Milan, 26-27-28 January 2022.

Chiarcos, C., Klimek, B., Fäth, C., Declerck, T., & McCrae, J. P. (2020, May). On the Linguistic Linked Open Data Infrastructure. In Proceedings of the 1st International Workshop on Language Technology Platforms (pp. 8-15).

Gracia J., Fäth C., Hartung M., Ionov, M., Bosque-Gil, J., Veríssimo, S., Chiarcos, C., Orlikowski, M. Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain. In: Fu B, Polleres A, eds. Proc. of 19th International Semantic Web Conference (ISWC 2020). Springer

References

Alva Principe, R. A., Maurino, A., Palmonari, M., Ciavotta, M., & Spahiu, B. (2021). ABSTAT-HD: a scalable tool for profiling very large knowledge graphs. *The VLDB Journal*, *1*(1), 26.

Debbatista, J., Auer, S., & Lange, C. (2016). Luzzu - A Methodology and Framework for Linked Data Quality Assessment. *Journal of Data and Information Quality*, *4*(1), 32.

Spahiu, B., Maurino, A., & Palmonari, M. (2018). *Towards Improving the Quality of Knowledge Graphs with Data-driven Ontology Patterns and SHACL*. ISWC (Best Workshop Papers), 103-117.