Report and Training Materials of the 2nd Training School

Project Acronym	NexusLinguarum
Project Title	European network for Web-centred linguistic data science
COST Action	CA18029
Starting Date	26 October 2019
Duration	48 months
Project Website	https://nexuslinguarum.eu
Responsible Authors	Jorge Gracia, Fahad Khan, Patricia Martín-Chozas
Contributors	Andon Tchechmedjiev, Christian Chiarcos
Version | Status	v2 | final
Date	September 30th, 2022

Acronyms List

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

Table of Contents

EXECUTIVE SUMMARY

This document reports on the 2nd training school organised by the NexusLinguarum COST Action. The training school was held physically from May 30th to June 3rd 2022 at Residencia Lucas Olazábal of Universidad Politécnica de Madrid, Cercedilla, Madrid. This training school was organised as a new iteration of the summer datathon series on Linguistic Linked Open Data, therefore constituting the *4th Summer Datathon on Linguistic Linked Open Data* (SD-LLOD-22). See https://datathon2022.linkeddata.es/ for more details.

The SD-LLOD-22 datathon had the main goal of giving people from industry and academia practical knowledge in the field of Linked Data applied to Linguistics. The final aim is 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.

The training school was attended by 39 participants: 26 of them were trainees, 5 organisers, 12 tutors and lecturers and 1 invited speaker.

In addition to the training activities, the characteristics and atmosphere of the event were also very conducive to scientific networking and a lively exchange of ideas between all types of participants, no matter if they were trainees, tutors or lecturers.

1. Introduction

The ultimate goal of the NexusLinguarum Action is to promote the study of linguistic data science, for which the construction of a mature holistic ecosystem of multilingual and semantically interoperable linguistic data is required at Web scale. Training schools are one of the means for reaching this goal, and therefore the NexusLinguarum team organised a second training school as a new iteration of the datathon series on Linguistic Linked Open Data (LLOD), the *4th Summer Datathon on Linguistic Linked Open Data* (SD-LLOD-22), which took place from May 30th to June 3rd 2022 at Residencia Lucas Olazábal of Universidad Politécnica de Madrid, Cercedilla, Madrid. See https://datathon2022.linkeddata.es/. This is the continuation of the first training school of NexusLinguarum, which took place on February 8-12, 2021 with the title "Introduction to Linked Data for Linguistics" as a purely online training school and that was organised under the umbrella of the EUROLAN series of training schools (see http://eurolan.info.uaic.ro/2021/index.html).

The 2nd training school (SD-LLOD-22) aimed at promoting and teaching the foundations of linguistic data science and its related technologies to people from both academia and industry. It brought together scholars, practitioners, teachers and students in relevant disciplines as linguistics, NLP and information technology to discuss principles and best practices for representing, publishing and linking linguistic data and issues that constitute building blocks in the envisioned multilingual and interoperable Web-oriented ecosystem.

Differently to the first NexusLinguarm training school, which was purely online, this was a purely physical meeting. That gave the participants the opportunity of a closer iteration and a much more favourable environment to develop a number of mini projects, which would be much more difficult in an online setting. In addition to the training activities, the characteristics and atmosphere of the event were also very conducive to scientific networking and a lively exchange of ideas between participants.

The school was organised and funded primarily by the NexusLinguarum COST Action (https://nexuslinguarum.eu/). However other projects contributed in a way or in another with their scientific support and/or covering the participation of some invited lecturers. These supporting projects were: Prêt-à-LLOD (https://pret-a-llod.github.io/), LiLa (https://lila-erc.eu/#page-top), KATY (https://katy-project.eu/#page-top), and MLY (https://www.ehu.eus/eu/web/mlv/home)

The training school was attended by 39 participants: 26 of them were trainees, 5 organisers, 12 tutors and lecturers and 1 invited speaker.

2. Topics and scope

The training school has been developed for newcomers as well as for those already having basic knowledge in the fields covered. The school provided a comprehensive introduction to methodologies for representing linguistic resources using Semantic Web technologies, together with means to extract knowledge from language resources and exploit it using Semantic Web query languages and reasoning capabilities. The topics addressed in the school were the following:

- Ontologies and Linked Data
- The Lexicon Model for Ontologies (Ontolex-Lemon)
- Integrating documents, annotations and NLP tools with Linked Data and RDF using Web Annotation and NIF
- Guidelines for RDF generation and publication of Language Resources

- Linked Data in lexicography and terminology
- Use and Applications of Linguistic Linked Data
- Metadata and Licenses for Linguistic Linked Data
- Linked Data-aware NLP workflows

During the datathon, participants acquired skills to:

- Generate and publish their own Linguistic Linked Data from some existing data sources.
- Apply Linked Data principles and semantic technologies (knowledge graphs, RDF, SPARQL) to the field of language resources.
- Use the principal models used for representing Linguistic Linked Data, in particular OntoLex lemon.
- Learn about Linked Data-based NLP workflows and applications.
- Learn about potential benefits and applications of Linguistic Linked Data for specific use cases.

3. Program and activities

The program of the summer datathon contained three types of sessions:
1. Seminars to show novel aspects and discuss selected topics.
2. Hands-on sessions to introduce the basic foundations of each topic, methods, and technologies and where participants will perform different tasks using the methods and technologies presented.
3. Datathon sessions, where participants worked, in groups of 3-4, on miniprojects and where they applied what they had learned, involving the generation and/or use of Linguistic Linked Data.

Participants were invited to propose a "**miniproject**" related to the topics of the datathon, which might include some datasets for their conversion into linked data. In this edition, we particularly encouraged miniprojects that involved **under-resourced languages**. The organisers made a selection and merging of the proposals, which formed the basis for the miniprojects which the participants could select to work on during the datathon sessions. Participants who did not propose a miniproject, or whose miniproject was not selected, were able to join another miniproject. An honorific diploma was awarded to the best miniproject.

The program had an initial common part for all the participants for the introduction of the more general topics. On the second day we introduced two tracks in parallel for the hands-on sessions: basic and intermediate (for SPARQL and LLOD generation). For the third and fourth day, we created two parallel tracks for the hands-on sessions and some seminars: i) corpora and annotation, and ii) lexicography and terminology, so participants could select one or another according to their interests. The "datathon" sessions were devoted to free work by the participant groups assisted by their tutors.

There was also a rich **social programme**, aimed at strengthening ties between participants and stimulating networking and scientific exchange in an informal setting. This comprised an *ice breaking* session on the arrival day with social games; a *hiking excursion* around the surrounding forest area of the Residence, and an *excursion to the city of Segovia*.

This was the program with the planned activities:

	Sun 29/5	Mon 30/5	Tue 31/5	Wed 1/6	Thu 2/6	Fri 3/6
09:00 - 09:30		Opening and Ontology/Linked Data basics	Presentation of participant groups	Parallel Seminars: Linguistic Annotations; Terminology and Lexicography	Seminar: Metadata	Seminar: LD-aware NLP workflows (Teanga)
09:30 - 10:00					Parallel Hands-on: Corpora Annotation;	
10:00 - 11:00		Seminar: Linguistic Linked Data and Ontolex	Seminar: SPARQL	Hands-on: Linking (NAISC and VocBench)	Lexicographical Data	Datathon: Results Presentations
11:00 - 11:30		Coffee Break	Coffee Break	Coffee Break	Coffee Break	Coffee Break
11:30 - 12:00		Introduction to VocBench and LiLa	Parallel Hands-on: Basic SPARQL Querying;	Parallel Hands-on: Corpora; Terminology and Lexicography	Datathon	Datathon: Results Presentations
12:00 - 12:30			Intermediate SPARQL Querying			Invited Talk: Artem Revenko
13:00 - 14:30		Lunch	Lunch	Lunch	Lunch	Lunch
14:30 - 15:00		Hands-on Ontolex Lexicon Building	Parallel Hands-on: Basic LLOD Generation;	Daily Report (tutors only)	Daily Report (tutors only)	Conclusions and Awards
15:00 - 15:30			Intermediate LLOD Generation	Datathon	Datathon	
16:00 - 16:30		Coffee Break	Coffee Break	Coffee Break	Coffee Break	
16:30 - 19:30	Arrival, Registration and Installfest	Minute Madness, Projects and Groups Selection, Datathon	Datathon	Excursion to Segovia	Datathon	
19:30 - 20:30			Hiking around Cercedilla			
20:30 - 22:00	Dinner and Icebreaking Session	Dinner	Dinner	Dinner	Dinner	

Figure 1. Program of the training school.

A PDF version of the program can be found here

4. Datathon participants

The training school was attended by 39 participants: 26 of them were trainees, 5 organisers, 12 tutors and lecturers and 1 invited speaker. The organisers were involved in the design of the course, the selection of lecturers, and the decision making in general; they also gave some lectures. The local organiser took care of the local logistics. The lecturers were invited to give one or several seminars or hands-on sessions. Some lecturers acted also as tutors, staying for the whole duration of the datathon and supervising the work done in the mini-projects. This is the list of organisers, lecturers, and tutors:

Organisers
==========
Jorge Gracia (University of Zaragoza, Spain)
Patricia Martín-Chozas (Universidad Politécnica de Madrid, Spain)

Anas Fahad Khan (Institute for Computational Linguistics «A. Zampolli»/CLARIN-IT, Italy)
Christian Chiarcos (Goethe Universität Frankfurt, Germany)

Local organiser
==========
Elena Montiel-Ponsoda (Universidad Politécnica de Madrid, Spain)

Tutors
==========
Thierry Declerck (DFKI, Germany)
Milan Dojchinovski (CTU in Prague, Czech Republic / DBpedia Association, Germany)
Dagmar Gromann (University of Vienna, Austria)
Max Ionov (Goethe Universität Frankfurt, Germany)
Christian Fäth (Goethe Universität Frankfurt, Germany)
David Lindemann (UPV/EHU University of the Basque Country, Spain)
Gilles Sérasset (Université Grenoble Alpes, France)
Andon Tchechmedjiev (IMT École des Mines d'Alès)

Lecturers
==========
Manuel Fiorelli (University of Rome Tor Vergata, Italy)
Francesco Mambrini (Università Cattolica del Sacro Cuore, Italy)
Bernardo Stearns (NUI Galway, Ireland)
Armando Stellato (University of Rome Tor Vergata, Italy)

Invited speaker
==========
Artem Revenko (Semantic Web Company)

As for the participant's profile, we circulated a survey after the school to gather that information (jointly with the general feedback of the datathon). Here are some results, showed graphically (figures 2 to 5):

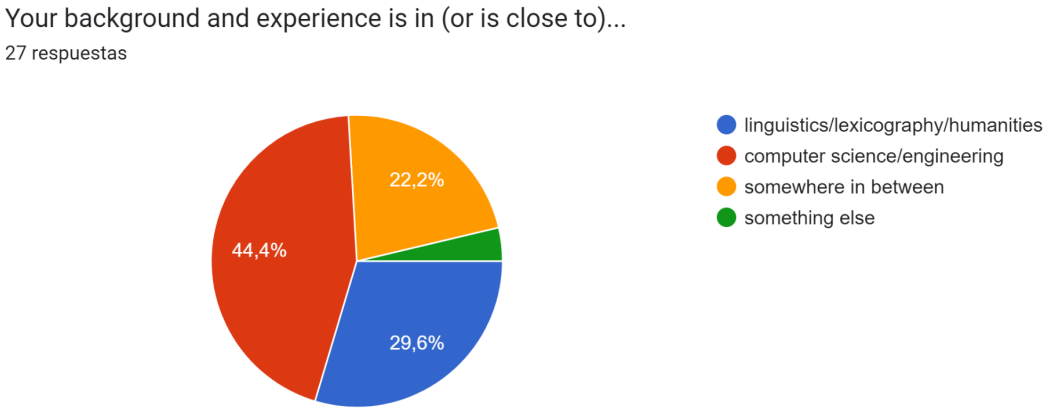

Figure 2. Participant's background

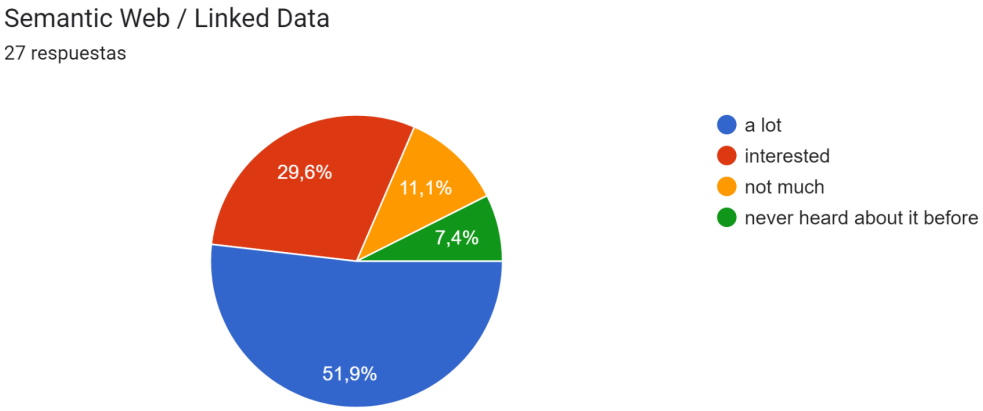

Figure 3. Participant's previous interest in Semantic Web / linked data

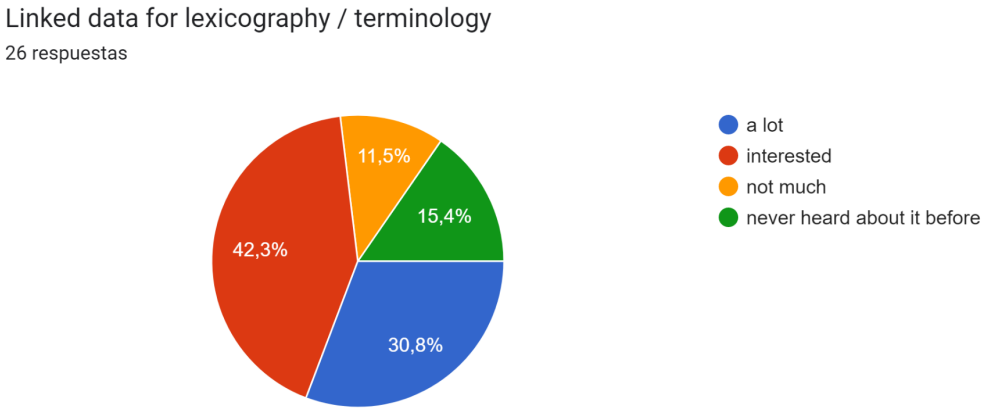

Figure 4. Participant's previous interest in lexicography / terminology

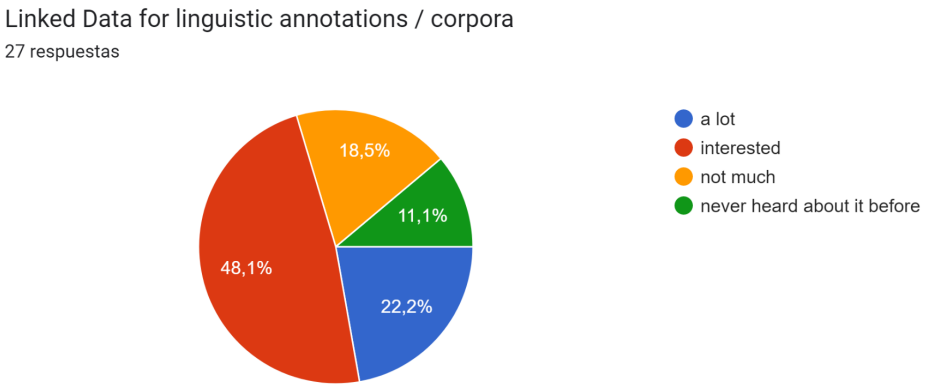

Figure 5. Participant's previous interest in linguistic annotation / corpora

5. Training Materials

Various types of materials have been generated for the training school, including presentations (slides) and exercises accompanied by code and data examples. All the materials were published in Zenodo and are made freely available:

Slides (DOI): https://doi.org/10.5281/zenodo.7197674

Practical Sessions' Materials (DOI): https://doi.org/10.5281/zenodo.7197696

6. Summary of the mini-projects

After the "groups formation" session during the first day, a total of 8 mini-project were finally selected to be developed during the datathon, each of them with an assigned tutor and a group of trainees to work on it. This is the list of mini-projects and participating group members:

Biomedical Terminology

Participants: Lucia Sanchez, Carlos Golvano and Renzo Alva Principe

Tutor: Dagmar Gromann

Description:

The aim of this project was to extract a biomedical terminology from a Spanish language corpus. To this end a bio-medical terminology extraction NLP pipeline was set up. This pipeline begins with a term extraction phase. In this case 3 different term extraction tools were compared for the task: Term Suite, TBXTools, and Text2TCS; with the latter of these emerging as a clear winner with 32066 terms extracted from the corpus. Next the pipeline features a post processing phase where the most frequent general Spanish words from the candidate terms are removed and the remaining terms are POS tagged with more terms being filtered

out on the basis of linguistic patterns. Finally in this pipeline the list of terms is enriched and published in SKOS.

Creating a Medical Termnet

Participants: Bettina Klimek, Erik Körner, Robert Lew

Tutor: Thierry Declerck

Description:

The project aimed to enrich a pre-existing medical dataset in order to enable transitive medical concept hierarchies as well as formalising semantic attributes and features of medical concepts and lexicalising basic medical concepts for discovering known and generating unknown complex medical concepts. The results of the project include the creation of a OntoLex lexicon, Medical Termnet; the creation of domain ontology with interrelated concepts. In addition the team investigated the application of string-match based NER tagging (using spaCy) over a medical corpus of 20,000 sentences. VocBench 3 was used for data editing and querying.

Dictionary LODification using Wikibase: Quechua language

[WINNER OF THE BEST MINI-PROJECT AWARD]

Participants: Valeria Caruso, Ibai Guillén, Elwin Huaman

Tutor: David Lindemann

Description:

The aim of the project was to create a LOD version of a dictionary for the endangered South American language Quechua on Wikibase. The project began with the identification of suitable resources. Next there was the preprocessing and setup stage. This latter included: setting up QICHWABASE, programming a bot as well as normalising, cleaning and refining the data identified in the first phase. A pre-set ontological Wikibase application profile was identified as a model and was subsequently populated. Finally the data was published and a SPARQL endpoint made available.

Lexis: A Computational Lexicon of Modern Greek

Participant: Katerina Gkirtzou

Tutor: Christian Faeth

Description:
The project had the aim of converting a computational lexicon for modern Greek currently stored in a relational database into linked data. Unfortunately the R2RML specifications lack the appropriate complex transformation mechanisms. The approach followed in the project began with an initial mapping of the

relational data to RDF vocabularies (in this case Ontolex family and lexInfo) to produce a set of draft triples. These were exported using SQL queries from the relational database. Next the Fintan tool was used to update these draft triples using a mapping mechanism. As a result of this 63,730 lexical entries were created and 206,742 triples overall.

LLOD Cloud Compass

Participants: Danella Gregg, Giedrė Valūnaitė Oleškevičienė

Tutor: Gilles Sérasset

Description:

The aim of the project was to assist users in determining the 'direction' to take in LLOD cloud when it comes to translating a dataset into another language. The resource designed by the team, the compass, helps users find which linked data resources are most suitable for their task by providing coverage and perplexity measures. The output of the project includes an Ontolex based version of the BATS dataset; the use of compass on two dataset samples (BATS and Rethfig); an experiment using four predefined translation strategies: DBnary direct translation, DBnary translation (with path length = 2), DBpedia, Wikidata.

Old English Metaphor Wordnet

Participants: Lucía Pitarch Ballesteros, Isam Diab, Rafael Cruz González

Tutor: Andon Tchechmedjiev

Description:

The aim of the project was to convert and enrich an input dataset describing etymological information about shame terms in the lexicon of Old English (77 lexical entries covering figurative expressions). The plan was to start by curating and standardising the data. Next the data was to be converted to Ontolex and subsequently linked to other resources. The team managed to carry out data analysis, data curation and standardisation. They also came up with a RDF-based model for modelling figurative meaning. The resulting RDF triples were mapped to EmotionOntology and MetaNet.

SENTiMiENTOS

Participants:Martín Alejandre Chaya, Miloš Košprdić, Tijana Radović

Tutor: Sina Ahmadi

Description:

This project aimed to look into a lexicon based approach for sentiment analysis of low-resource languages. It started off with pre-prepared sentiment lexicons for Serbian and German containing lemmas and corresponding sentiment scores. During the project these lexicons were enriched with PoS tags and senses and converted into RDF triples using Marl, Ontolex, Lexinfo vocabularies. In addition the team created a

web-based sentiment analysis application on the basis of the data which they had enriched for showing the relative polarity of Serbian and German words.

To OntoLex

Participants: Zaida Bartolome, Khadija Ait Elfqih, Kristina Kocijan

Tutor: Max Ionov

Description:

The aim of the project was to convert the prototype of a bilingual French-Spanish terminology dictionary for the architectural domain to linked data using Ontolex (hence the name of the project). The original dataset consisted of entries extracted from two corpora built from texts found on the internet and validated by a group of experts. These entries were accompanied by their translation, definitions and contexts. In order to carry out the conversion the team made use of the OpenRefine tool as well as producing a template script that can potentially be adapted to any XML serialisation. Vocbench 3 was also used for editing entries. One outcome of this experience of working on this project was a positive evaluation of Vocbench as a very useful tool with a clear pedagogical function

5. Conclusions

The second NexusLinguarum training school, organised as a new iteration of the Summer Datathon on Linguistic Linked Data, provided valuable knowledge and trained a number of computer scientists and linguists on how to work and benefit from linguistic linked data. This edition continued the first online training school that took place in February 2021. Such first school was introductory to the topic and put the basis for this second edition, which has progressed towards more advanced and specialised topics around two driving themes: corpora and linguistic annotation on the one hand and lexicography and terminology on the other. An eye was also put on LD-aware NLP workflows. Emergent user-friendly tools such as VocBench were extensively treated in the school as well as successful use cases such as the LiLa project.

All the materials created during the training school are publicly available and can be further used and utilised by the community.

After the feedback survey we could confirm that the level of satisfaction with the school was very high in general. The majority of participants felt that they had benefited a lot from all the types of sessions as well as from the social activities. They felt that they had acquired new knowledge in all the treated topics. Most of the participants considered that the datathon's atmosphere was conductive to learning (90%) and to networking (100%). As for suggestions for improvements, they mentioned longer datathon (free work) sessions, more social activities, and more basic introduction into semantic web techniques.

Appendix - Some pictures of the event

Figure 2. A hands-on session in terminology.

Figure 3. Participants working on their miniprojects.

Figure 4. Excursion to Segovia

Figure 5. Group picture