Report and Training Materials of the 1st 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	Milan Dojchinovski and Julia Bosque-Gil
Contributors	Jorge Gracia, Sara Carvalho, Ilan Kernerman, Thierry Declerck
Version | Status	v2.2 | final
Publication Date	29/04/2021 (updated 27/10/22)

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 1st training school organized by the NexusLinguarum COST Action. 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 modeling and querying linguistic data with state-of-the-art ontology models and tools. The training school has been organized 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.

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 core team organized an Introduction to Linked Data for Linguistics online training school that took 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. It was organized under the umbrella of the EUROLAN series of Summer Schools, which has been established in 1993 and covers topics that are particularly relevant to the fields of computational linguistics and natural language processing (NLP). The goal of this 15th EUROLAN School was to bring 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.

2. Scope and Program

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:

- Semantic Web and linked data
- Ontologies (RDF, RDF-S, OWL, etc.)
- Query mechanisms (SPARQL)
- Metadata (DCAT, VOID, etc.)
- RDF transformation and validation
- Linguistic linked data
- Lemon-Ontolex
- Linguistic linked data generation
- Corpora and linked data
- Linguistic annotations
- Tools and applications of linguistic linked data

Summary of the training school:

- The first day started with an opening session and a brief introduction to Linguistic Linked Data (LLD), followed by an introduction to Linked Data and RDF dedicated sessions.
- The second day covered topics related to ontologies, including modelling knowledge with ontologies, OWL and SKOS knowledge representation languages, reasoning of knowledge, and a hands-on session using the ontology editor Protégé.
- The third day was dedicated to topics regarding representation and querying lexical data with dedicated sessions on the Ontolex-Lemon model and the SPARQL querying language.

- The fourth day included sessions which gave an overview of other linguistic and metadata vocabularies and the VocBench platform modeling linguistic datasets. In the late afternoon, the local organizers organized an online social event where the participants could closely, although remotely, “taste” the beauty of the Romanian culture, traditions and nature.
- The fifth day comprised three parallel sessions on different topics: (i) LLD Generation/Transformation and Linking, (ii) Annotations (NIF, Web Annotation), and (iii) Ontolex Extensions (*vartrans*, *lexicog*, FrAC). Finally, the training school ended with a closing session in which an ontology of participants, lecturers and organisers was presented and which illustrated many of the representation mechanisms explained throughout the week (see Figure 1 in the appendix).

Each of the organized sessions was accompanied by a hands-on session and an exercise session. During the hands-on session, the lecturers proposed an exercise and offered a step-by-step walk-through for the participants to understand the methodology towards its solution. They also introduced the basic technology needed. Then, during the exercise session, the participants were asked to work on a particular task similar to the cases presented during the hands-on session, thereby becoming familiar with the technology introduced in a practical setting. As these sessions were arranged in terms of complexity, starting with the basic notions and building on to present more specific topics in a detailed fashion on the last day, participants had the chance to acquire a solid foundation before moving onto more complex sessions.

The official program of the school is available online and included here as Appendix I.

3. Organization and Logistics

Due to the COVID-19 pandemic and current travel restrictions in Europe and beyond, the training school was held online. EUROLAN 2021 has been virtually hosted 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. Tutorials coupled with hands-on sessions have been held throughout the week, 8-12 February, 2021. The training school has been driven and organized by WG1 members of the NexusLinguarum COST Action. Twelve lecturers were involved in the organization.

Following on the EUROLAN tradition of over almost three decades, which is known for the excellence of the academic programs along with camaraderie among professors and students, a range of virtual activities were arranged in addition to the online classes, with the aim of providing cultural experiences and discoveries, in addition to closer interaction. Attendance was online and free of charge, requiring pre-registration. The organizing committee consisted of:

- Jorge Gracia, University of Zaragoza, Spain
- Christian Chiarcos, Goethe-University Frankfurt am Main, Germany
- Milan Dojchinovski, CTU in Prague, Czech Republic / InfAI at Leipzig University, Germany
- Daniela Gîfu, "Alexandru Ioan Cuza" University of Iasi & Romanian Academy – Iași Branch
- John McCrae, National University of Ireland, Galway, Ireland
- Eric Curea, Romanian Academy, Bucharest, Romania

The organization was supported by a steering committee, as follows:

- Dan Cristea, Romanian Academy – Iași Branch, Romania
- Daniela Gifu, University of Iași (UAIC) & Romanian Academy - Iasi Branch, Romania
- Jorge Gracia, University of Zaragoza, Spain
- Nancy Ide, Vassar College, New York, US
- Dan Tufiș, Romanian Academy, Bucharest, Romania

All the sessions were hosted using the Zoom platform. For the hands-on sessions, several breakout (virtual) rooms were enabled where the participants could work on the assignment in smaller groups. To encourage participants to ask questions and get in touch with each other, the organizers set up a Slack channel where lecturers and participants would clarify any doubts. The total number of participants was 82, 52 female and 30 male.

4. 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.7258979`

Practical Sessions' Materials (DOI): `https://doi.org/10.5281/zenodo.7258991`

5. Summary

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 organized by the NexusLinguarum COST Action 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.

During the closing session (photos presented in Appendix II), the organizers provided participants with a survey form (see Appendix III) to gather feedback on both organisational and academic aspects of the school, which was completed by over 25% of the attendees. The results show that the disciplines of humanities/linguistics/lexicography had a higher representation among participants than computer science, and that the school was considered to be well focused, well balanced topic-wise and well organised. In particular, theory sessions, tutoring, and the opportunities to learn were very well evaluated. On the other hand due to the virtual mode, there is still room for improvement in practical sessions, social event organisation and opportunities to network.

Appendix I

EUROLAN-2021 Program

Day1 (8/2/21) - Linked Data Basics

Morning (times in CET)

09:00 -09:50 **Session 1**: Welcome and Introduction; Overview to Linguistic Linked Data

Jorge Gracia,Dan Tufiș, Dan Cristea, Daniela Gîfu

09:50 -10:00 Break

10:00 -10:50 **Session 2**: Linked Data Principles, RDF, RDF-S

Thierry Declerck, Julia Bosque-Gil

10:50 -11:00 Break

11:00 -12:00 **Hands-on session**: RDF

Thierry Declerck, Julia Bosque-Gil

Afternoon (times in CET)

13:30 -16:00 **Practical exercises (free work)**: RDF

Thierry Declerck, Julia Bosque-Gil, Jorge Gracia

16:00 -17:00 Exercises results

Day2 (9/2/21) - Ontologies

Morning (times in CET)

09:00 - 09:50 **Session 1**: Modeling Knowledge with Ontologies

Thierry Declerck, Julia Bosque-Gil

09:50 -10:00 Break

10:00 -10:50 **Session 2**: OWL, SKOS, Basic Reasoning

Thierry Declerck, Julia Bosque-Gil

10:50 -11:00 Break

11:00 -12:00 **Hands-on session**: Ontology Edition in Protégé

Sina Ahmadi

Afternoon (times in CET)

13:30 -16:00 **Practical exercises (free work)**: Ontologies

Sina Ahmadi, Max Ionov, Christian Chiarcos, Jorge Gracia

16:00 -17:00 **Exercises results**

Day3 (10/2/21) - Representing and querying lexical data

Morning (times in CET)

09:00 - 09:50 **Session 1**: Linguistic Linked Data and OntoLex-Lemon

John McCrae

09:50 -10:00 Break

10:00 -10:50 **Session 2**: Querying Semantic Data (SPARQL)

ChristianChiarcos, Max Ionov

10:50 -11:00 Break

11:00 -12:00 **Hands-on session**: SPARQL

Christian Chiarcos, Max Ionov

Afternoon (times in CET)

13:30 -16:00 **Practical exercises (free work)**: SPARQL over Linguistic Data

Christian Chiarcos, Max Ionov, Andrea Turbati

16:00 -17:00 **Exercises results**

Day4 (11/2/21) - Other vocabularies + VocBench

Morning (times in CET)

09:00 -09:50 **Session 1**: Other Linguistic and Metadata Vocabularies (DCat, Lexinfo, ...)

John McCrae

09:50 -10:00 Break

10:00 -10:50 **Session 2**: Introduction to the VocBench Platform

ArmandoStellato, ManuelFiorelli

10:50 -11:00 Break

11:00 -12:00 **Hands-on session**: A Guided Tour to VocBench

Armando Stelatto, ManuelFiorelli

Afternoon (times in CET)

13:30 -16:00 **Practical exercises (free work)**: VocBench for modeling linguistic datasets

Armando Stelatto, ManuelFiorelli, AndreaTurbati

16:00 -17:00 **Exercises results**

19:30 -21:00 **Online Social Event**

Dan Cristea, Daniela Gîfu

Day5 (12/2/21) - Advanced topics (parallel sessions)

Parallel session 1: Linguistic Linked Data Pipeline

09:00 -09:50 **LLD Generation/Transformation**

Max Ionov

09:50 -10:00 Break

10:00 -10:50 **Linking LLD**

Sina Ahmadi

Parallel session 2: Annotations

09:00 -09:50 **Linguistic Annotations (WA, NIF)**

Christian Chiarcos

09:50 -10:00 Break

10:00 -10:50 **NLP Web services (NIF)**

Milan Dojchinovski

Parallel session 3: Ontolex Extensions

09:00 -09:35 **Variation and translation (vartrans)**

Jorge Gracia

09:35 -10:10 **Lexicographical Data (lexicog)**

Julia Bosque-Gil

10:10 -10:20 Break

10:20 -10:50 **Frequency, Attestation, and Corpus (frac)**

Christian Chiarcos

11:00 -12:00 Closing session

Nancy Ide, Jorge Gracia, Dan Tufiș, Daniela Gîfu

Appendix II

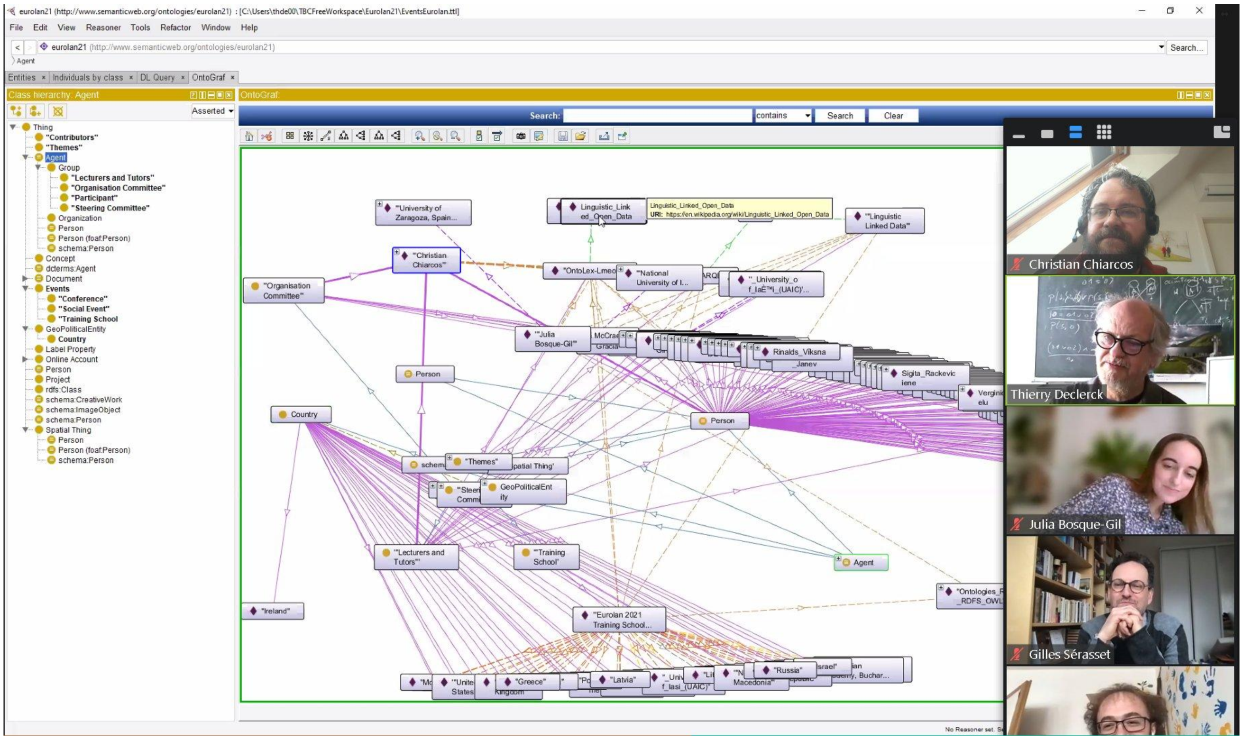

Figure 1. Ontology of the training school.

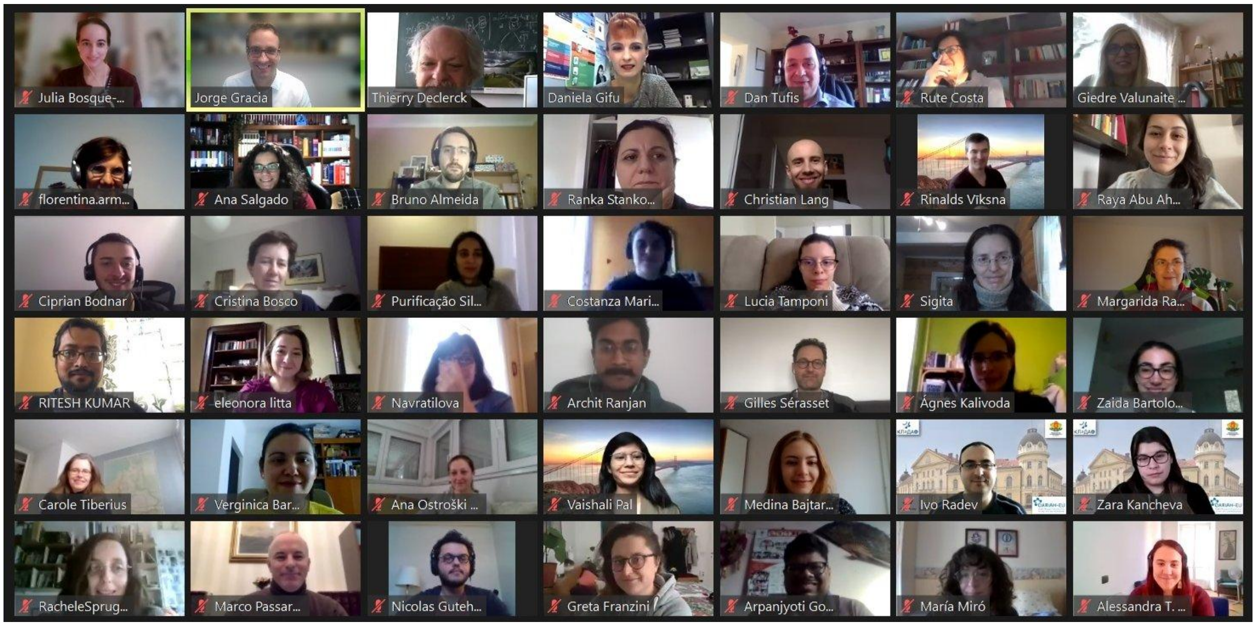

Figure 2. Group photo with a number of participants.

Appendix III

(attachment on the next page)