MOOC on Linguistic Linked Data

Jorge Gracia[1], Slavko Žitnik[2], Max Ionov[3], Christian Chiarcos[4], Dagmar Gromann[5], Francesco Mambrini[6], Marco Passarotti[6], Armando Stellato[7], John P. McCrae[8], Gilles Sérasset[9], Andon Tchechmedjiev[10], Sara Carvalho[11], Penny Labropoulou[12], and Rute Costa[13]

[1] University of Zaragoza, Spain jogracia@unizar.es
[2] University of Ljubljana, Slovenia slavko.zitnik@fri.uni-lj.si
[3] University of Cologne, Germany max.ionov@gmail.com
[4] Universität Augsburg, Germany christian.chiarcos@uni-a.de
[5] University of Vienna, Austria dagmar.gromann@univie.ac.at
[6] Università Cattolica del Sacro Cuore, Italy
{francesco.mambrini,marco.passarotti}@unicatt.it
[7] University of Rome Tor Vergata, Italy stellato@uniroma2.it
[8] University of Galway, Ireland john.mccrae@nuigalway.ie
[9] Université Grenoble Alpes, France gilles.serasset@imag.fr
[10] Institut Mines-Telecom, France andon.tchechmedjiev@mines-ales.fr
[11] University of Aveiro, Portugal sara.carvalho@ua.pt
[12] Athena Research Center, Greece penny@athenarc.gr
[13] Universidade Nova de Lisboa rute.costa@fcsh.unl.pt

Abstract. The field of linguistic linked data lies at the intersection of the Semantic Web and linguistics and studies techniques and tools aimed at modelling and publishing language resources on the Web, in ways that enable their data interoperability and reuse. In order to support learning in this area and train a new generation of researchers and practitioners, the NexusLinguarum COST Action developed a MOOC on linguistic linked data. This freely accessible MOOC is unique in its kind and has been prepared by experts with broad experience in the field. During the course, students acquire fundamental concepts of linguistic linked data and gain practical experience with related tools and techniques. The MOOC consists of two courses: (i) *essentials*, which covers basic tools and modelling techniques (e.g., Ontolex-Lemon, SPARQL, VocBench, NIF) and (ii) *advanced topics*, covering more advanced contents like metadata, lexicography, terminology, deep learning and linguistic data, and a real use-case. The courses are self-paced and the expected duration of both is seven weeks. We expect the course to have a strong impact by promoting the combination of Semantic Web technologies with linguistics and data science, which is crucial in fields of emerging studies such as *linguistic data science*. The community behind this initiative is well-anchored and coordinates multiple recent standards for linguistic knowledge representation increasingly adopted by linguists.

Keywords: linguistic linked data · MOOC · linguistic data science

1 Introduction

The linked data (LD) paradigm emerged years ago as a set of best practices and principles to expose, share, and connect data on the Web in a semantically interoperable manner [3]. More than a decade ago, a research community emerged that studied how such LD principles could be applied to the modelling of linguistic data and language resources (dictionaries, terminologies, corpora, etc.), considering the peculiarities of application domains such as corpus linguistics, computational linguistics, and natural language processing (NLP). The field that studies the LD paradigm when applied to the specificities of linguistic data is known as *linguistic linked data* (LLD) [9,8]. The community behind such a field has been very active in developing vocabularies, best practices, and tools, but also in understanding the benefits of the LD approach as well as systematizing the field, bridging the gap between the advances in linguistics and language technologies and those taking place in the Semantic Web and artificial intelligence areas [16,17].

In this context, the *European network for Web-centred linguistic data science* (NexusLinguarum)[14] COST Action was aimed at building an ecosystem of Web-scale multilingual and semantically interoperable linguistic data technologies and resources. An essential ingredient in building such an ecosystem is education on its related technologies (notably LLD) to newcomers in the area. In fact, the entry point to LLD for non-experts is typically perceived as not straightforward and requires substantial time and effort to acquire the concepts. To facilitate this, a number of initiatives have been developed during the last years such as textbooks, conference tutorials, and training schools (see Section 2). However, the elaboration of a MOOC (massive open online course) was still pending. The NexusLinguarum community supported this idea and materialised it, resulting in the educational resource that we report in this paper: the first MOOC on LLD[15], which was made available online on 25th September 2024.

The MOOC on LLD complements other existing training materials and initiatives, but adds several advantages: (i) accessibility: It is available to anyone with an internet connection, thus breaking geographical and financial barriers, (ii) flexibility: Learners can access the materials at any time and study at their own pace, and (iii) variety: It includes videos, slides, readings, and interactive elements such as quizzes, which enhance the learning experience. During the MOOC on LLD, the learner acquires the fundamental notions around LLD and gains practical experience with its main tools and techniques.

The remainder of this paper is organised as follows. In Section 2 some related learning initiatives are mentioned. Then, the structure and organisation of the MOOC is summarised in Section 3 and a closer look at its contents is provided in Section 4. The elaboration methodology is explained in Section 5 and some insights on the potential impact of the MOOC can be found in Section 6. Section 7 contains the concluding remarks.

[14] `https://nexuslinguarum.eu/`

[15] `https://nexuslinguarum.eu/results/mooc/`.

2 Related resources and initiatives

At the time of writing, there is no other MOOC on the topic of LLD. There are, however, other online courses on closely related topics such as knowledge graphs, Semantic Web, and linked data. Such courses are more general and constitute an excellent complement to the MOOC on LLD[16], serving to settle its grounds. However, they are not sufficient to cover the specificities of linguistic data and applications in a Semantic Web context.

We mention, for instance, the MOOC on "Semantic Web Technologies" at openHPI, the educational Internet platform of the German Hasso Plattner Institute, Potsdam[17], which was first launched in February 2013 and updated in later editions[18]. The course introduces the fundamentals of Semantic Web technologies, and how to represent and access semantic data on the Web. The course is free and self-paced, although initially scheduled for six weeks. Such a course was followed by a good number of other related courses at openHPI, such as: "Knowledge Engineering with Semantic Web Technologies"[19], "Linked Data Engineering"[20]. "Knowledge Graphs"[21] and "Knowledge Graphs - Foundations and Applications"[22].

There are also MOOCs on "Ontology Engineering" and "Semantic Web and Linked Data" by Universidad Politécnica de Madrid (Spain), first launched in 2017 through the MiriadX platform, with several later editions[23], as well as the course "Mastering ontology development: a practical approach"[24] since 2021. The typical duration of such courses is six weeks. Other popular courses are "Semantic web with python"[25], at Udemy platform, and the four-week course on "Introduction to a Web of Linked Data" at FUN (France Université Numérique)[26].

Focusing on the topic of LLD, a good number of initiatives (tutorials, datathons, training schools) have had a remarkable influence in the area and contributed to training, disseminating, and raising awareness. In fact, part of the materials used for this MOOC was initially elaborated and validated in some of these events, particularly in the Summer Datathon on Linguistic Linked Open Data (SD-LLOD) series. Such an initiative started in June 2015 in Cercedilla (Spain)

[16] We are not counting here the plethora of videos and tutorials that can be found on the Web about ontologies and Semantic Web, but only courses that were built and structured as online courses and are available through any MOOC platform.

[17] https://open.hpi.de/courses/semanticweb

[18] e.g., https://staging.openhpi.de/courses/semanticweb2017

[19] https://staging.openhpi.de/courses/semanticweb2015

[20] https://open.hpi.de/courses/semanticweb2016

[21] https://open.hpi.de/courses/knowledgegraphs2020

[22] https://open.hpi.de/courses/knowledgegraphs2023

[23] https://eventos.upm.es/127085/detail/semantic-web-and-linked-data-edicion-17.html

[24] https://eventos.upm.es/125928/detail/mastering-ontology-development-a-practical-approach-edition-8.html

[25] https://www.udemy.com/course/introduction-the-semantic-web-with-python/

[26] https://www.fun-mooc.fr/en/courses/introduction-web-linked-data/

and celebrated its fifth edition in June 2023 in Lužnica (Croatia)[27]. The slides and training materials were made available online after each datathon edition. We also mention here the 2015 and 2021 editions of the international EUROLAN summer school[28], both of them devoted to the topic of LLD.

Finally, we mention the first monographic book on LLD, titled "Linguistic Linked Data. Representation, Generation and Applications" [9], published by Springer in 2020 and co-authored by three of the MOOC on LLD lecturers. This book presents a mix of background information on linguistic linked data and concrete implementation advice, and describes in detail how linked data principles can be applied to modelling and exploiting language resources. The book has been very influential in the elaboration of the MOOC and is an important complementary reference.

3 Structure and organisation of the MOOC

In this section, we summarise the main features of the MOOC on LLD and give an overview of its structure and content organization.

3.1 Main features

- *Pre-requirements.* In principle, the only pre-requirement to follow this course is a basic understanding of the main techniques and standards of the Semantic Web (i.e., OWL ontologies, RDF, RDFS, and the basic principles of Linked Data). Students do not need to be experts on it, but at least be familiarised with the basic notions of the Semantic Web. In any case, one of the first lessons of our course is a quick overview of Linked Data and the Semantic Web, to refresh these basic concepts. The lesson also provides pointers to external courses and materials that can be used in preparation for this MOOC.
- *Target audience.* The target group of this MOOC is everyone interested in language technologies, willing to represent and generate linguistic data in standard and interoperable ways on the Web.
- *Duration.* The estimated *duration* of the MOOC is seven weeks with 3-4 hours of dedication per course week. However, the MOOC is self-paced, thus the duration is just indicative and can be adjusted to the student's needs.
- *Qualification.* There are a number of quizzes and assignments embedded in the lessons. They serve for *self-assessment* and progress checking. However, they are not supervised and are not intended to issue any diploma or qualification certificate.

[27] https://datathon2023.jezik.hr/

[28] See https://conferences.info.uaic.ro/eurolan/2021/ and https://conferences.info.uaic.ro/eurolan/2015/

- *Publication* The MOOC on LLD is free and publicly available under a CC-BY-SA 4.0 license. It was published by the German University of Digital Science (UDS)[29] and divided into two courses, both of them accessible through the following web address[30]:

  `https://nexuslinguarum.eu/results/mooc/`

3.2 Structure

The MOOC is divided into two parts or courses: *essentials* and *advanced*. The essentials course introduces the basic notions of LLD and its foundational models such as Ontolex-Lemon, as well as some practical tools and frameworks to deal with LLD (representing, querying, publishing, etc.). The advanced course introduces some popular LLD resources (such as DBnary[31] and Wikidata[32]) and the application of LLD into specific fields (e.g., lexicography, terminology), as well as current topics such as deep learning in LLD. The lesson that wraps up the course describes the LiLa project, a network of interconnected linguistic resources (dictionaries, glossaries, corpora, etc.) for the study of the Latin language that exemplifies very well many of the concepts explained along the two parts of the MOOC, in a real setup.

Tables 1 and 2 outline the MOOC syllabus for both the essentials and advanced courses. Most of the lectures consist of the following materials:

1. Lesson: one or several videos and their corresponding slide files.
2. Assignment: assignment instructions with a separate solution document.
3. Quiz: questions to test the acquired skills and knowledge.
4. Reading materials: recommended complementary readings.

4 Contents and learning outcomes

In this section, we explain the main contents and learning outcomes of each of the lectures of the MOOC. We are skipping week 1 since their lessons are merely preparatory.

4.1 Essentials Course.

Linguistic Linked Data. This lesson introduces the concept of linguistic linked data, emphasising its potential to address challenges in integrating diverse language resources, such as dictionaries, corpora, and lexical resources, which often

[29] https://german-uds.academy/

[30] And also searchable through the UDS website. DOIs have also been created for them: 10.6084/m9.figshare.28596197 and 10.6084/m9.figshare.28596575 respectively.

[31] `https://kaiko.getalp.org/about-dbnary`

[32] `https://www.wikidata.org/`

week	**lecture title**	**short description**
1	Welcome and introduction	Welcoming words and brief introduction to the course
	Installing the tools	Installation of the tools used later in the course
	Semantic Web and linked data	Introduction to the SW and LD basic concepts (URIs, RDF, OWL, etc.)
2	Linguistic linked data	High-level overview over LLD vocabularies, community and resources
	Tools for ontolex lexicon building: VocBench	Showing existing lexicons and how to build a lexicon from scratch in VocBench
	Modelling: Ontolex-Lemon	The *lemon* model: modelling principles, modules, and examples
3	Linked data tools overview	Explanation of some LD tools and frameworks like Jena and Protégé
	SPARQL	Explain the SPARQL query language with hands-on examples
4	Linked data generation	LLD RDF generation from linguistic data in various formats using VocBench and the RDF Mapping Language (RML)
	Corpora and annotation	Different approaches and vocabularies for corpora annotation (e.g., NIF, Web annotation, CoNLL-RDF, OLiA) with practical hands-on examples

Table 1. Syllabus of the 'essential' course

week	**lecture title**	**short description**
5	Metadata	Models for metadata, general ones and specific of language data and resources
	Linked data resources	Big Linguistic Linked (Open) Data resources, such as Wikidata and DBnary
	Linked data and lexicography	Use of Ontolex-Lemon and its lexicog module to model lexicographic resources
6	Linked data and terminology	Overview to terminological data structures, terminology extraction, enrichment and representation in the Semantic Web
	Deep learning and linked data	Brief introduction to deep learning and neural language models in the context of linguistic data; e.g., neural relation acquisition and extraction, triple generation, and NL2SPARQL
7	Use case: LiLa project	Linking Latin (LiLa) project as a success story of linguistic linked data

Table 2. Syllabus of the 'advanced' course

exist in isolated, heterogeneous formats. It highlights the limitations of traditional approaches and presents RDF graphs as a solution to achieve interoperability, scalability, and FAIR (Findable, Accessible, Interoperable, and Reusable) principles. Key topics include the role of RDF in creating interoperable data structures, the integration of resources using shared vocabularies like Ontolex-Lemon and SKOS [7], and tools such as Recogito for annotation. The lesson underscores the importance of linking datasets and aligning metadata to enhance linguistic research and applications in areas such as lexicography, corpus linguistics, and natural language processing.

Tools for Ontolex Lexicon Building. In this lesson, the VocBench platform [26], a collaborative editing environment for modelling ontologies, knowledge organization systems (classification systems, authority tables, thesauri, etc..), lexicons and generic datasets is first introduced and then showcased in its specific support for the management of LLD. The VocBench lesson is articulated in three parts:

1. Introduction to the VocBench platform.
2. Hands-on session on how to develop an OWL ontology with VocBench.
3. Hands-on session on how to develop an Ontolex lexicon with VocBench and link it to an existing ontology. The lesson also showcases the use of Ontolex-Lemon Design Patterns through their dedicated implementation as VocBench Custom Forms [12].

Modelling: Ontolex-Lemon. This section covers modelling lexical resources with Ontolex-Lemon [10,19], one of the foundational models of LLD, as well as the SKOS model. Despite Lemon being more expressive, SKOS is also extensively used to model language resources, particularly thesauri, and it is a usual entry point to the Semantic Web in language technologies. The course covers the design of the model and outlines the design decisions that guided the creation of Ontolex-Lemon. We then describe the core modules of the model by means of examples both visually and in Turtle code. Then, a brief tour of the other modules for syntax and semantics, decomposition, variation, translation, and metadata is provided. Then, the SKOS model is described, which can be used as an alternative to OWL as a target for Ontolex-Lemon models. This provides a basic introduction to the usage of both Ontolex-Lemon and SKOS models.

Linked data tools overview. The goal of this lesson is to guide students through using some of the essential tools that they will need later in the course and in practice to work with RDF datasets and ontologies (to explore, edit, and query them). It shows how to explore ontologies and RDF vocabularies with the ontology editor Protégé,[33] which is crucial for understanding the data the students will encounter, both in the course and later, if they choose to apply the skills gained in this MOOC.

[33] `https://protege.stanford.edu/`

The next part of the lesson concerns writing RDF data from scratch in Turtle, a simple and human-readable RDF serialization. Having done this, students are then guided through adding Turtle data to Apache Jena Fuseki graph database.[34] that can later be queried with SPARQL.

SPARQL. This lesson introduces linked data querying, especially focused on non-technical students. It explains the fundamentals of SPARQL, the query language designed to access RDF data across the Web. The lesson explores how SPARQL enables sophisticated queries on RDF graphs, dives into its syntax and capabilities, and applies it through practical examples using real-world databases like DBpedia or Wikidata. The lecture is organised around a general querying example related to automobiles. The participants build their learning on top of the example and get to know SPARQL concepts while querying DBpedia.

Programming (querying) assignment includes repetition of all the queries from the lecture against the live DBpedia SPARQL endpoint, and adapting all the queries to get *same* results from Wikidata. Lastly, participants are asked to query linguistic concepts in Wikidata and then explore on their own.

Linked Data Generation. The objective of the linked data generation lesson is to allow students of the course to automate the generation of linked data resources in a principled way. Given the broad demographic of students with both linguists and more technically inclined profiles, we made the choice of offering two tracks: one using the high-level user interface of VocBench for LLD generation, based on the integrated Sheet2RDF tool [25], the other using the lower-level RDF Mapping Language (RML [11][35]) using the RML engine from the `morph-kgc` python library. A first video capsule introduces LD generation, followed by specific video capsules for VocBench and RML that the students could choose from. Subsequently, a practical use case to generate LLD from Wiktionary data is proposed as a project that can be approached with either of the two tools. A comprehensive solution and accompanying notebooks are provided, as well as a series of quizzes for each video capsule that allow students to self-assess their ability.

Corpora and Annotation. This lesson touches on three different topics. The first is to define what a linguistic corpus actually is. To that end, some sample corpora in the CoNLL format are examined, along with its Linked Data representation with CoNLL-RDF and showing how to query that data with SPARQL. The second part covers Web Services, and this is where LD technology really generates a benefit over conventional corpus technology. Here, the learner can

[34] `https://jena.apache.org/documentation/fuseki2/`

[35] We consider the unofficial specification available at `https://rml.io/specs/rml/`, as it is the only one currently supported in implementations. A new version of the standard is in preparation [13], but since the new version is back-compatible with the unofficial specification, only minor adaptations would be required to the material.

explore how to use a remote SPARQL endpoint for integrating information from a lexical knowledge graph. As a second example, entity linking with DBpedia Spotlight [21] is explored. In the third part, we briefly tackle a few slightly more advanced aspects of corpora and annotation. Accompanying materials are also provided, with additional literature and references to the systems and technologies introduced here.

4.2 Advanced Topics Course.

Metadata. This lesson aims at giving all the tools to students to assist them in exploiting metadata for the discovery and usage of resources, as well as in generating principled metadata for their LLD resources. The lesson starts with an introduction to basic metadata concepts, metadata types, their objectives, and their representation in LD formats, all clarified with real examples and use cases. It goes on to cover best practices for the management and standardization of data and metadata, especially recommendations formulated in the FAIR principles [28], that students should take into account when publishing their LLD resources. Some popular catalogues and repositories hosting data and services are included in the lesson. Finally, the lesson takes a deeper dive into general-purpose metadata models (e.g., DCAT [1], VoID [2]) and two models that cater specifically for the description of linguistic data and services, namely META-SHARE [17,20], and lime, which is the metadata module of Ontolex [10]. The lesson concludes with two step-by-step screencasts that show students how to describe their datasets and data processing services, using as an example a service presented in the terminology lesson and its output dataset. A comprehensive metadata template for datasets, combining elements from the models presented in the lesson, is adapted to the example resources. Students are provided with the template and encouraged to use it for their LLD resources prior to submission to linked data portals and aggregators.

Linked Data Resources. Structured in five parts, this lesson presents the two largest LLD datasets freely available on the Internet: DBnary [24] and Wikidata [27]. For each of these resources, their characteristics (size, structure, specific approaches. . .) are introduced. Then, the lesson addresses the way in which one may use them directly through their public website and how one may easily browse the dataset, query their SPARQL endpoint, or download and mirror the data locally. Then, each resource is explored by querying it with many diverse queries that are explained in natural language and implemented as SPARQL queries that gradually rise in complexity.

At the end of this lesson, the student should be able to query each of these resources (and many others available that are represented using the Ontolex-Lemon model) through SPARQL queries that they may use to prototype dictionary access to tenth of millions of lexical entries in their own use case.

Linked Data and Lexicography. This lesson begins by explaining why linked data and a graph-based representation are useful for lexicographic data. Next, different models of representation of lexicographic data as linked data are reviewed and, finally, a real example of a lexicographic resource as linked data is shown.

At the end of the lesson, the student will have acquired and worked on:

1. The value of links and graph-based structures in lexicographic data [4].
2. Examples of lexicographic resources represented as linked data.
3. The use of Ontolex vartrans and Ontolex lexicog [5] modules to represent lexicographic data[36].
4. Example of a lexical resource converted into linked data: the K-Dictionaries use case [6].

Linked Data and Terminology. This lesson aims to explore how terminological data can be represented and shared as Linked Data, resorting to examples from the medical domain. It is structured into four main parts: (1) understanding terminology and its key concepts, (2) extracting terminology, (3) enriching terminology, and (4) integrating ontologies into terminology work. The first section introduces terminology as both a collection of terms and a scientific discipline, following ISO 1087:2019 [15] principles, and focuses on the double dimension (linguistic and conceptual) of terminology. Other key concepts addressed in this section include terms, concepts, and concept relations. The second section delves into terminology extraction from unstructured texts, exploring a tool such as Text2TCS[37], which extracts terms, groups them by synonymy, and finds relations between them. The tool is available on the European Language Grid[38]. The third section addresses terminology enrichment and introduces TermitUp [18], a terminology management tool that automates the linking of terms to external LD resources such as Wikidata and IATE[39]. The final section focuses on the integration of ontologies into terminology work, and their role in organising concepts systematically and ensuring data interoperability. Tools such as Protégé enable the representation of more complex concept systems, paving the way to more effective knowledge sharing across multilingual and multidisciplinary contexts.

Deep Learning and Linked Data. Structured in four main parts, this lesson provides a general overview of deep learning with a particular focus on neural language models and methods for querying, obtaining, and verbalizing Linked Data. First, a general overview of neural language models and different methods of (pre-)training is offered. Secondly, the lesson connects to previous lessons on SPARQL by introducing current approaches to automatically learn structured queries from natural language text. Thirdly, approaches to directly extract knowledge graphs and acquire relations from natural language text in multiple languages are exemplified. Finally, to conclude the roundtrip, neural approaches

[36] https://www.w3.org/2019/09/lexicog/

[37] https://text2tcs.univie.ac.at/en/

[38] https://live.european-language-grid.eu/

[39] https://iate.europa.eu/home

that verbalize triples in order to obtain natural language text from structured data, such as Knowledge Graphs and Linked Data, are introduced. The contents of videos, quizzes, and accompanying assignments are complemented with references to allow participants to further explore these topics.

Use case: LiLa Project. This final lesson describes the LiLa: Linking Latin project[40], a successful use case in LLD, which exemplifies many of the concepts explained along the course, in a real setup. LiLa [23] is an LLD-driven Knowledge Base of interconnected language resources (like dictionaries, glossaries, corpora, etc.) for the study of the Latin language. The core of LiLa consists of a large collection of more than 200k canonical forms of citation of Latin words, called the Lemma Bank: interoperability is achieved by linking all those entries in lexical resources and tokens in corpora that point to the same lemma. The lesson is divided into five parts, covering the following learning objectives:

1. The architecture of the LiLa Knowledge Base.
2. Modeling and interlinking lexical resources in LiLa.
3. Modeling and interlinking textual resources in LiLa.
4. The SPARQL Endpoint of the LiLa Knowledge Base[41].
5. The online services to query and populate LiLa, with practical exercises.

5 Elaboration and publication.

The MOOC on LLD has been elaborated in the context of the NexusLinguarum COST Action. The topics and materials were developed in synergy with other Action's activities such as training schools, mobility grants, and the preparations of an Erasmus Mundus master programme on Linguistic Data Science. The instructional design of the MOOC followed broadly accepted practices for complex learning and associated instructional methods [22].

There have been altogether 23 contributors to the published MOOC version, one of them acting as coordinator and 13 as lecturers[42]. Lecturers were responsible for preparing and giving the recorded lessons. The MOOC elaboration group had 17 online meetings, along with several on-site encounters during the NexuLinguarum plenary meetings, and one on-site meeting in Athens for the video recordings. The whole process of elaboration of the course was as follows:

1. **Related courses and publishing platforms review.** We found eleven related MOOCs that covered only partial aspects of our proposed MOOC. We thoroughly examined their syllabi, durations (specific lecture videos and as a whole), accompanying materials, and activities (quizzes, assignments).

[40] `https://lila-erc.eu`. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme – Grant Agreement No. 769994.

[41] `https://lila-erc.eu/sparql/`

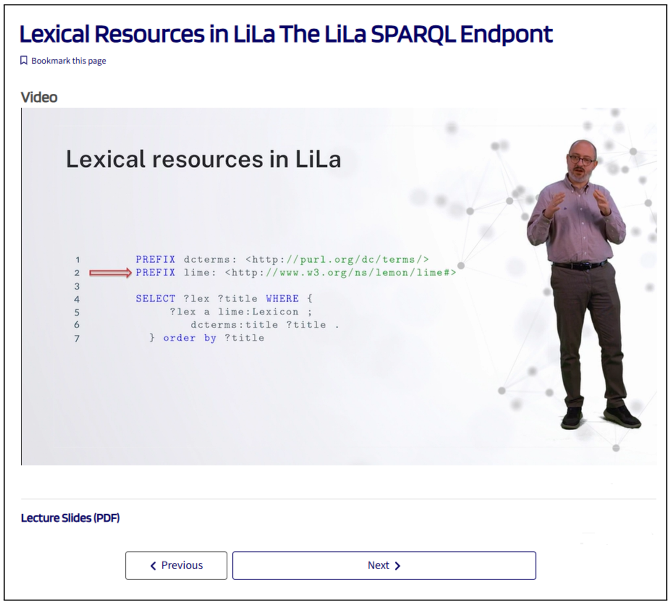
[42] This paper is co-authored by such lecturers.

In this step, we also reviewed nine promising MOOC platforms. Especially, we searched for platforms that allow for free publishing options, offer free courses without much supervision, and already have an appropriate audience. Accessibility considerations are also important in a MOOC design [14] and the selected platform should cover most of them well, such as responsiveness, compatibility across multiple devices, and consistency in design. The openHPI platform [43] seemed the most appropriate. At the time our MOOC was under preparation, part of the openHPI group was starting a new edX-based platform - German UDS, and we agreed to publish the MOOC there.

2. **Broad topics definition.** We defined the main topics of the MOOC. The topics were related to data science in linguistics, although we only briefly covered general topics (that are also part of other courses) so that the MOOC can be self-sufficient. We planned to develop a seven-week course, so the division of amount of time per topic was decided at this step.
3. **Lessons specialization.** Each topic consists of one or more lessons. Each lesson comprises one or several video recordings, slides, and accompanying material – quizzes, reading assignments, or programming assignments. We specialized the topic syllabus with the lesson data, a short description of lectures and their outcomes, and also the prerequisites needed, expected duration, and accompanying materials definition. At this step, we had identified 16 different lessons. Larger lessons were split into multiple recordings.
4. **Contributors selection.** We invited other experts in the field to contribute to the MOOC. For each lesson, we selected a coordinator (lecturer), who prepared, with the assistance of other collaborators, all the materials and presented the lesson during the video recording session.
5. **Material preparation instructions.** We defined a detailed MOOC preparation plan with all the deadlines until the on-site recording meeting in Athens. The lecturers were instructed to prepare their materials in various formats and they would be unified during the post-editing step. The most important material was a transcript of each lecture. The transcripts allowed to precisely define the video lengths and to check the coherence of the contents. We also updated the transcripts and added intros and outros, so that all the lectures were connected. Each lecturer was also instructed to practise their recordings so that on-site recordings would be more efficient. For the quizzes, the lecturers were instructed to prepare single or multiple-choice questions (with solutions) only. Programming/reading assignments should be in text form with optional code/other data in a zip format.
6. **On-site recordings.**
   An instructional designer assisted us in reorganising the transcripts and provided advice about what to wear, how to stand, how fast to talk, etc. We had a very limited time for recordings as all the recordings were shot in two consecutive days. Immediately after the recordings, the storyboarding process was completed - i.e., selection of shot type (full-size lecturer shot, *cowboy*-shot, slide-shot only), placement of slide vs. lecturer on the video

[43] `https://open.hpi.de`

(left or right side) – see Fig. 1. At that time, we also gathered the lecturers' profile photos, institution names, and their short descriptions.



Lexical Resources in LiLa The LiLa SPARQL Endpont

Lexical resources in LiLa

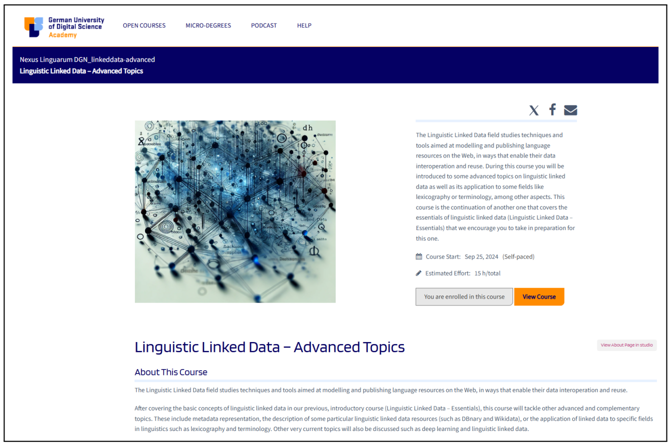
```
      PREFIX dcterms: <http://purl.org/dc/terms/>
      PREFIX lime: <http://www.w3.org/ns/lemon/lime#>

      SELECT ?lex ?title WHERE {
           ?lex a lime:Lexicon ;
              dcterms:title ?title .
        } order by ?title
```

Fig. 1. Example of video lesson in the UDS portal

7. **Post-editing.** Recording and video post-editing were performed by personnel of the Institute for Language and Speech Processing in Athens. At this time, we started unifying all the materials, which was time-consuming owing to the different slide templates used by the lecturers. Common templates were also created for the assignments. A short description and all the MOOC metadata were defined also in this phase, prior to its publication.
8. **Courses publishing.** Courses were published in mid-September 2024 [44]. Each course is freely available to the public. The landing page of each course (Fig. 2) contains an overview of the course and expected workload.

   The MOOC platform stores the course participant's progress, so each person can clearly see their progress and follow each course at their own pace (see Fig. 3). For better organisation, courses are organised by weeks and by topics within every week.

[44] https://german-uds.academy/courses/course-v1:NexusLinguarum+DGN_linkeddata-essentials+2024_1, https://german-uds.academy/courses/course-v1:NexusLinguarum+DGN_linkeddata-advanced+2024_1, accessed March 15, 2025.



Nexus Linguarum DGN_linkeddata-advanced

Linguistic Linked Data – Advanced Topics

The Linguistic Linked Data field studies techniques and tools aimed at modelling and publishing language resources on the Web, in ways that enable their data interoperation and reuse. During this course you will be introduced to some advanced topics on linguistic linked data as well as its application to some fields like lexicography or terminology, among other aspects. This course is the continuation of another one that covers the essentials of linguistic linked data (Linguistic Linked Data – Essentials) that we encourage you to take in preparation for this one.

Course Start: Sep 25, 2024 (Self-paced)

Estimated Effort: 15 h/total

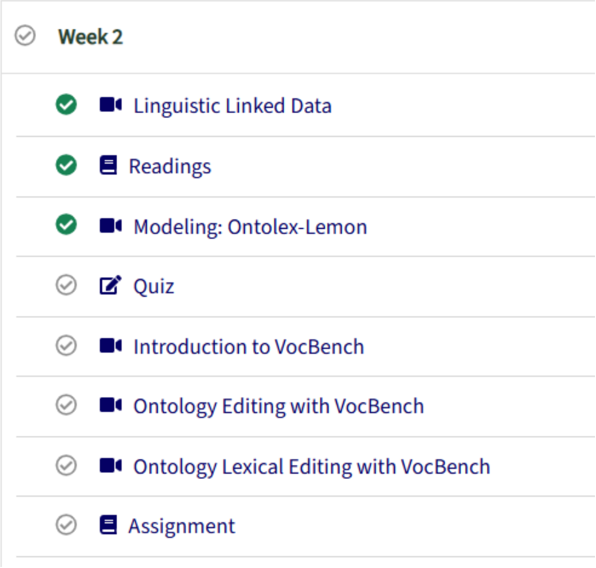
Linguistic Linked Data – Advanced Topics

About This Course

The Linguistic Linked Data field studies techniques and tools aimed at modelling and publishing language resources on the Web, in ways that enable their data interoperation and reuse.

After covering the basic concepts of linguistic linked data in our previous, introductory course (Linguistic Linked Data – Essentials), this course will tackle other advanced and complementary topics. These include metadata representation, the description of some particular linguistic linked data resources (such as DBnary and Wikidata), or the application of linked data to specific fields in linguistics such as lexicography and terminology. Other very current topics will also be discussed such as deep learning and linguistic linked data.

Fig. 2. Landing page of the advanced course



Week 2

- ✓ Linguistic Linked Data
- ✓ Readings
- ✓ Modeling: Ontolex-Lemon
- Quiz
- Introduction to VocBench
- Ontology Editing with VocBench
- Ontology Lexical Editing with VocBench
- Assignment

Fig. 3. Example of lesson organisation (week 2) as it appears in the UDS portal. Green circles denote elements completed by the learner.

6 Impact

During its first 80 days, the number of enrolled students was 108 for the 'essentials' course and 42 for the 'advanced topics' course. Given its recent publication, it is too early to measure the real impact of the MOOC in the community. However, these early numbers show a good initial uptake, which is expected to grow after future dissemination campaigns.

On the other hand, the future impact of the MOOC is granted by the recently approved European Master in Linguistic Data Science (EMLDS)[45]. Such a two-year international master is also an important outcome of the NexusLinguarum Action and is aimed at training a new generation of researchers and practitioners in the emergent field of linguistic data science. The first edition of the master will begin in 2025-26. The MOOC on LLD will be used as preparatory material for the master's students, in particular before the third semester, in which they will be specifically trained on LLD.

We expect that its open licence scheme, its availability in a consolidated learning platform, and its free aspect will favour the incorporation of the MOOC's learning materials in the training programmes of different universities and institutions, particularly (but not exclusively) those that were part of the NexusLinguarum Action.

7 Conclusions

In this paper, we have presented the MOOC on LLD, recently published via the German UDS. It is structured around two parts or courses: essentials and advanced, to better fit the learner's interest and background. This freely accessible MOOC is unique in its kind and has been prepared by experts with broad technical and pedagogical experience in the field. The MOOC allows students to acquire fundamental concepts of linguistic linked data and gain practical experience with related tools and techniques. The course has great potential for linguists and computer scientists willing to learn about how to represent and share linguistic data on the Web in a formal and interoperable way. Its initial uptake is promising and its impact will be further amplified by the expected use of these materials in the future European Master on Linguistic Data Science and, potentially, in many other training initiatives.

Acknowledgments. We are very grateful to Panagiotis Karioris for his valuable contribution to the video recordings and post-edition, and to Gokhan Ozkan for his help with the recording preparation and organisation. This article is based upon work from COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It has been also partially supported by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE), by DGA/FEDER, and by the *Agencia Estatal de Investigación* of the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the "Ramón y Cajal" program (RYC2019-028112-I).

[45] `https://emlds.fcsh.unl.pt/`

References

1. Data catalog vocabulary (DCAT) - version 3, https://www.w3.org/TR/vocab-dcat-3/, w3C Working Draft
2. Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J.: Describing linked datasets with the VoID vocabulary, https://www.w3.org/TR/void/
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems **5**, 1–22 (7 2009). https://doi.org/10.4018/jswis.2009081901
4. Bosque-Gil, J., Gracia, J., Gómez-Pérez, A.: Linked data in lexicography. Kernerman Dictionary News pp. 19–24 (7 2016)
5. Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E.: Towards a module for lexicography in ontolex. In: Proc. of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets at 1st Language Data and Knowledge conference (LDK 2017), Galway, Ireland. vol. 1899, pp. 74–84. CEUR-WS (6 2017)
6. Bosque-Gil, J., Lonke, D., Gracia, J., Kernerman, I.: Validating the ontolex-lemon lexicography module with k dictionaries' multilingual data. In: Proc. of 6th biennial conference on electronic lexicography, eLex 2019. pp. 726–746. Lexical Computing CZ s.r.o. (2019)
7. Brickley, D., Miles, A.: SKOS core guide. W3C working draft, W3C (Nov 2005), https://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
8. Chiarcos, C., Nordhoff, S., Hellmann, S. (eds.): Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata. Springer (2012), http://dblp.uni-trier.de/db/books/daglib/0028690.html
9. Cimiano, P., Chiarcos, C., McCrae, J.P., Gracia, J.: Linguistic Linked Data. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-30225-2
10. Cimiano, P., McCrae, J.P., Buitelaar, P.: Lexicon Model for Ontologies: Community Report (2016), https://www.w3.org/2016/05/ontolex/
11. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: Bizer, C., Heath, T., Auer, S., Berners-Lee, T. (eds.) Proceedings of the 7th Workshop on Linked Data on the Web. CEUR Workshop Proceedings, vol. 1184 (Apr 2014), http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf
12. Fiorelli, M., Lorenzetti, T., Pazienza, M.T., Stellato, A.: Assessing vocbench custom forms in supporting editing of lemon datasets. In: Gracia, J., Bond, F., McCrae, J.P., Buitelaar, P., Chiarcos, C., Hellmann, S. (eds.) Language, Data, and Knowledge. pp. 237–252. Springer International Publishing, Cham (2017), https://link.springer.com/chapter/10.1007%2F978-3-319-59888-8_21
13. Iglesias-Molina, A., Van Assche, D., Arenas-Guerrero, J., De Meester, B., Debruyne, C., Jozashoori, S., Maria, P., Michel, F., Chaves-Fraga, D., Dimou, A.: The rml ontology: A community-driven modular redesign after a decade of experience in mapping heterogeneous data to rdf. In: Payne, T.R., Presutti, V., Qi, G., Poveda-Villalón, M., Stoilos, G., Hollink, L., Kaoudi, Z., Cheng, G., Li, J. (eds.) The Semantic Web – ISWC 2023. pp. 152–175. Springer Nature Switzerland, Cham (2023)
14. Iniesto, F., Rodrigo, C.: Understanding accessibility in moocs: Findings and recommendations for future designs. Journal of Interactive Media in Education (2024), https://api.semanticscholar.org/CorpusID:272816164

15. ISO: 1087:2019 Terminology Work and Terminology Science — Vocabulary. International Organization for Standardization, Geneva (2019)
16. Khan, A.F., Chiarcos, C., Declerck, T., di Buono, M.P., Dojchinovski, M., Gracia, J., Oleskeviciene, G.V., Gifu, D.: A survey of guidelines and best practices for the generation, interlinking, publication, and validation of linguistic linked data. In: Proc. of the 8th Workshop on Linked Data in Linguistics (LDL 2022) at LREC 2022. pp. 69–77. ELRA (2022), https://aclanthology.org/2022.ldl-1.9.pdf
17. Khan, A.F., Chiarcos, C., Declerck, T., Gifu, D., García, E.G.B., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J.P., Émilie Pagé-Perron, Passarotti, M., Muñoz, S.R., Truică, C.O.: When linguistics meets web technologies. recent advances in modelling linguistic linked data. Semantic Web **13**, 987–1050 (9 2022). https://doi.org/10.3233/SW-222859
18. Martín Chozas, P., Vázquez Flores, K., Calleja Ibáñez, P., Montiel Ponsoda, E., Rodríguez Doncel, V.: Termitup: generation and enrichment of linked terminologies. Semantic Web **13**(6), 967–986 (September 2022). https://doi.org/10.3233/SW-222885
19. McCrae, J., Spohr, D., Cimiano, P.: Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In: Proc. of the 8th Extended Semantic Web Conference. pp. 245–249 (2011). https://doi.org/10.1007/978-3-319-59888-8_17
20. McCrae, J.P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., Cimiano, P.: One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the web. In: Gandon, F., Guéret, C., Villata, S., Breslin, J., Faron-Zucker, C., Zimmermann, A. (eds.) The Semantic Web: ESWC 2015 Satellite Events, pp. 271–282. Lecture Notes in Computer Science, Springer International Publishing (2015), https://link.springer.com/chapter/10.1007/978-3-319-25639-9_42
21. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: Dbpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems. p. 1–8. I-Semantics '11, Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2063518.2063519
22. van Merriënboer, J.J.G., Clark, R.E., de Croock, M.: Blueprints for complex learning: The 4c/id-model. Educational Technology Research and Development **50**, 39–61 (2002), https://api.semanticscholar.org/CorpusID:11197753
23. Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F.M., Litta, E., Moretti, G., Ruffolo, P., Sprugnoli, R.: Interlinking through lemmas. the lexical collection of the lila knowledge base of linguistic resources for latin. Studi e Saggi Linguistici **58**(1), 177–212 (2020)
24. Sérasset, G.: Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. Semantic Web **6**(4), 355–361 (2015)
25. Stellato, A., Fiorelli, M., Lorenzetti, T., Turbati, A.: Lifting spreadsheets made easy: Combining power and magic in sheet2rdf. In: Garoufallou, E., Sartori, F. (eds.) Metadata and Semantic Research. pp. 135–146. Springer Nature Switzerland, Cham (2024)
26. Stellato, A., Fiorelli, M., Turbati, A., Lorenzetti, T., Gemert, W., Dechandon, D., Laaboudi-Spoiden, C., Gerencsér, A., Waniart, A., Costetchi, E., Keizer, J.: Vocbench 3: A collaborative semantic web editor for ontologies, thesauri and lexicons. Semantic Web pp. 1–27 (05 2020). https://doi.org/10.3233/SW-200370, https://content.iospress.com/articles/semantic-web/sw200370
27. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)

28. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**, 160018 (Mar 2016). `https://doi.org/10.1038/sdata.2016.18`, `http://www.nature.com/articles/sdata201618`