# Linguistic linked open data and semantic web

## Prof. Francesco Mambrini

***COURSE AIMS AND INTENDED LEARNING OUTCOMES***

The course will provide an introduction to the principles and practice of working with Linked Open Data (LOD) in the field of linguistics. By the end of the course, the students will have gained a comprehensive understanding of the principles of LOD, the main technologies of the Semantic Web Architecture (RDF, SPARQL, OWL) and of the ontologies that are used to publish linguistic resources as linked data. They will acquire the technical skills necessary to build, query, and extract information from linked data resources. They will be able to apply this knowledge to enhance language technology applications, evaluate existing resources, and develop their own projects in this exciting and rapidly evolving field.

***COURSE CONTENT***

The course is articulated in the following sections.

1. Introduction to Linguistic Linked Data:
	1. The history: from the “web of documents” to the “web of data”.
	2. The Semantic Web: when, why and how it was conceived.
2. The Semantic Web technologies: an introductory overview.
	1. Representing data in RDF.
	2. Querying data with SPARQL.
	3. Modeling the data: vocabularies, schemas, and ontologies.
	4. Building ontologies with OWL.
3. Representing Linguistic Information as Linked Data:
	1. Lexical resources: the Ontolex-Lemon model.
	2. The Ontolex extensions: translation, semantics, morphology and beyond.
	3. Corpora and annotation.
4. Applications of Linguistic Linked Data:
	1. Review of successful linguistic linked data projects.
	2. The language and the “world”: linguistic linked data and Wikidata.

***READING LIST***

[All the readings reported below are not mandatory for the exam]

Cimiano, Ph., Ch. Chiarcos, J.P. McCrae, and Jorge Gracia. 2020. *Linguistic Linked Data: Representation, Generation and Applications*. Cham: Springer. <https://doi.org/10.1007/978-3-030-30225-2>.

Khan, A.F., et al. 2022. “When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Data.” *Semantic Web* 13 (6): 987–1050. <https://doi.org/10.3233/SW-222859>.

Passarotti, M., F. Mambrini, G. Franzini, F.M. Cecchini, E. Litta, G. Moretti, P. Ruffolo, and R. Sprugnoli. 2020. “Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin.” *Studi e Saggi Linguistici* 58: 177–212. <https://doi.org/10.4454/ssl.v58i1.277>.

***TEACHING METHOD***

The course will consist of a combination of lectures, hands-on exercises that will be performed together by the lecturer and the students, and a final project. In addition to traditional classroom instruction, students will have the opportunity to work with real-world linguistic linked data resources and to develop their own linked data project for the final exam.

***ASSESSMENT METHOD AND CRITERIA***

Oral exam, articulated in two parts:

discussion of a final project related to linguistic linked data;

questions and answers related to the topics discussed in class.

The final project will consist in the creation of a small linguistic resource (dataset, ontology, or application) as linked data. Some examples will be discussed in class. The students will be requested to submit the data and/or code of their project *before* the exam; during the exam, they will have to give a brief presentation of their work and be prepared to answer questions about it.

The second part of the exams aims to verify the degree of acquisition of the course contents and will consist of general questions on the topics discussed in class.

Both sections will contribute to the final grade, in the following proportion: 67% will be determined by the final project, 33% by the answers to the general questions about the course.

***NOTES AND PREREQUISITES***

The course requires some basic familiarity with the problems and methods of Computational Linguistics, as well as with linguistic resources. Students should also know how to install and execute open-source applications and command-line tools.

*Place and time of consultation hours*

By appointment (to schedule, write to: francesco.mambrini@unicatt.it).

Place: CIRCSE Research Center, Franciscanum building, second floor.