14:40 - 12:30 Demonstrations of the UPSKILLS Learning Content Blocks

Introduction to Language Data: Standards and Repositories

Iulianna van der Lek

Darja Fišer

With contributions from:

Francesca Frontini

Alexander Konig

Willem Elbers

Learning outcomes

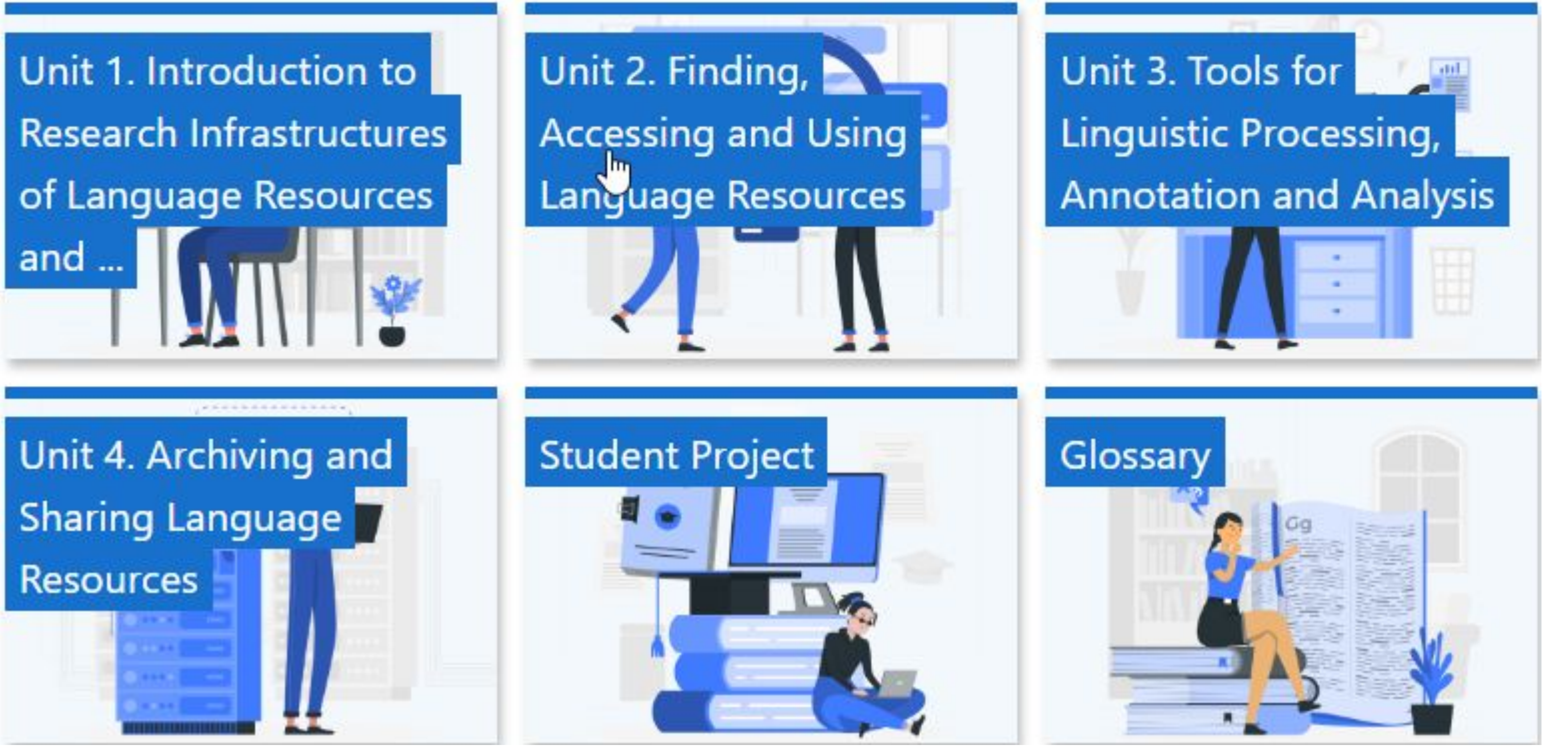
4-5 ECTS

By the end of this unit block, students will be able to:

- Explain what a language resource is and the role that research infrastructures play in the research data lifecycle in the context of Open Science and FAIR
- Use certified research data repositories to search, find and access language resources and datasets
- Process, annotate, and analyse different types of corpora in online environments according to standards and formats used by the community
- Archive and share language resources.

Prerequisites: Introduction to Text Processing (UniBo)

Course structure in Moodle



Unit 1. Introduction to Research Infrastructures of Language Resources and ...

Unit 2. Finding, Accessing and Using Language Resources

Unit 3. Tools for Linguistic Processing, Annotation and Analysis

Unit 4. Archiving and Sharing Language Resources

Student Project

Glossary

Highlights

- Learn by doing
- Interactive content slides in H5P and learning activities
- Take-home assignments and resources for self-study
- Modular: lessons can be picked and combined

Example of assignment

Search for 5 corpora in the CLARIN Resource Families on a topic that interest you and assess their FAIRness by answering the questions below:

- Findability: Are the corpora findable via Google/Bing, VLO and OLAC?
- Accessibility: Is the data accessible?
- Interoperability: In which format is the data available?
- Reusability: Is there documentation available on formats, methods and licensing?
- Other: Is the data openly available, is there a corpus paper or a dedicated website available?

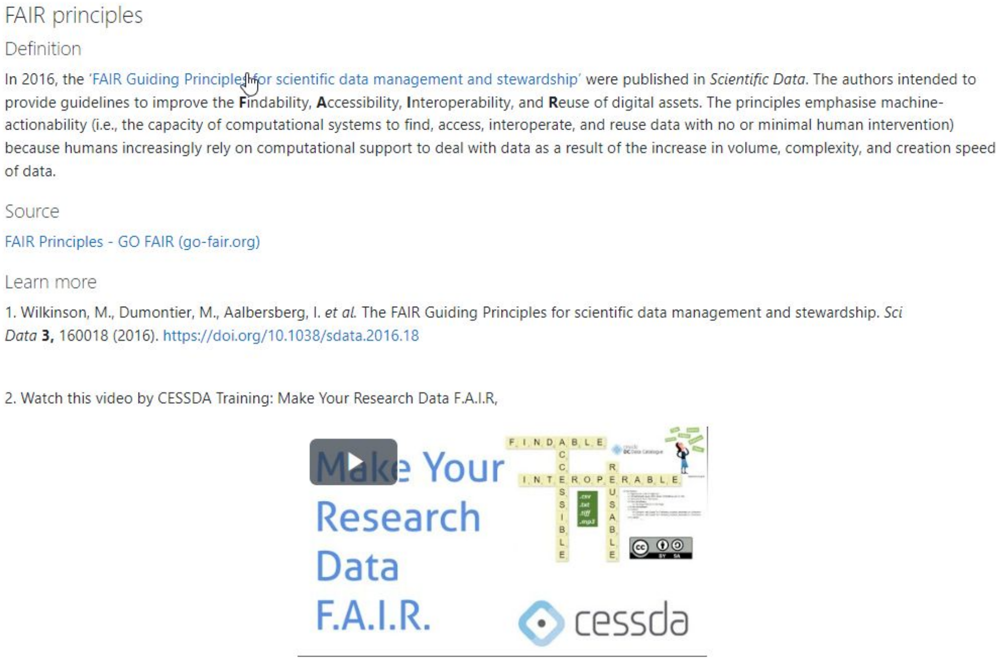
Delivery format: Blog post (800 words max).

Learning activity based on:

Frey, J.-C., König, A., & Stemle, E. W. (2019). How FAIR are CMC Corpora? Proceedings of the 7th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora2019), 25–30. https://cmccorpora19.sciencesconf.org/resource/page/id/15.

Glossary

Key concepts related to repositories, standards and research infrastructures



FAIR principles

Definition

In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in *Scientific Data*. The authors intended to provide guidelines to improve the **F**indability, **A**ccessibility, **I**nteroperability, and **R**euse of digital assets. The principles emphasise machine-actionability (i.e., the capacity of computational systems to find, access, interoperate, and reuse data with no or minimal human intervention) because humans increasingly rely on computational support to deal with data as a result of the increase in volume, complexity, and creation speed of data.

Source

FAIR Principles - GO FAIR (go-fair.org)

Learn more

1. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

2. Watch this video by CESSDA Training: Make Your Research Data F.A.I.R,