



Co-funded by the  
Erasmus+ Programme  
of the European Union



# Integrating Research Infrastructures into Teaching

By Iulianna van der Lek and Darja Fišer,  
CLARIN ERIC

CLARIN



Utrecht Multiplier Event, 4 November 2022

# Context and Motivation

## **UPSKILLS needs analysis:**

- Data and services offered by **research infrastructures** (e.g. CLARIN) are seldom used in teaching of language-related disciplines due to a general lack of knowledge
- Skills and knowledge about **linguistic data standards and repositories** were not explicitly mentioned in the learning outcomes of the analysed European language and linguistics degrees

# Context and Motivation

## **Survey of lecturers in language-related programmes (May-July 2021)**

- To learn about current practices in research and industry-based teaching methods, use of language resources and research data repositories
- 93 respondents, mainly from Linguistics, Language Teaching, Language and/or speech technologies, Translation and Interpreting, Linguistic Data Management

# Context and Motivation

## Research Data Discovery

- Little usage of repositories to find published language resources to use in teaching
  - institutional, CLARIN national repository, Linguistic Data Consortium, OPPUS, Corpora Mailing List, Language Resource Families, DGT Translation Memories, ELRA, Meta-Share

## Research Data Discovery

- Challenges:
  - Technical
  - Usability
  - Accessibility
  - Financial
  - Learnability

# Context and Motivation

## **Storage, archiving and sharing practices**

- Stored language resources on a shared cloud folder, e.g. Dropbox, Google drive
- Kept resources on their own computers
- Stored resources in their institutional repository

## **Repositories used for archiving resources**

- Github
- CLARIN national repository, subject-specific repository
- Meta-share
- ELRA Catalogue of Language Resources
- Language Data Consortium
- Zenodo, FigShare
- Moodle
- California Language Archive

# Context and Motivation

## **Storage, archiving, and sharing language resources**

- Technical challenges, limited space, little IT support
- Administrative load and costs
- Issues with IPR, students need help with the interpretations of the legal requirements
- Protection of data privacy in the case of spoken language and multilingual recordings
- Students' low level of digital literacy

# Why use RIs into your teaching?

## Advantages for students

- By interacting with research infrastructures (e.g CLARIN, DARIAH, Meta-Share), students gain awareness of **open science and FAIR** and may engage in collaborative research
- Opens new career paths, e.g. **language data manager** (clean, curate and manage data) or **data steward**

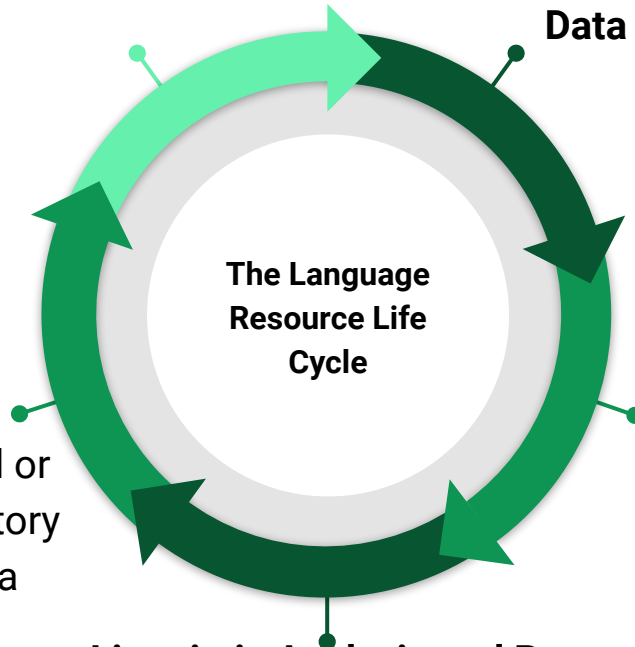
# Why use RIs in your teaching?

## Data Sharing

- Repositories
- Catalogues
- Citation
- Legal Issues

## Data Archiving

- Deposit LRs in a institutional or domain-specific FAIR repository
- Describe LRs using metadata standards
- IPR and legal issues for handling sensitive data; licenses



## Data Acquisition and Collection

- Research Infrastructures
- Repositories
- Data Catalogues
- IPR & Legal Issues
- Citation

## Data Curation and Annotation

- Standards and formats
- Annotation methods

## Linguistic Analysis and Research

- Querying metadata
- Analysis & visualisation; combining data from different sources -> exchange standards



# Filling the gap

1. Include specific RI-related learning outcomes in the **RBT guide**
2. **Quick guide** to CLARIN and how to use the main services and find relevant information
3. **Accompanying learning content:** *Introduction to Language Data: Standards and Repositories*
4. Close collaboration with UniBo to integrate the infrastructure into **the students' projects**
  - a. How to search and query existing corpora
  - b. How to archive corpora in a repository

# How to Use the CLARIN Research Infrastructure A Guide for Teachers and Trainers

## Authors:

- Iulianna van der Lek
- Darja Fišer

## Valuable contributions from the community:

- Francesca Frontini
  - Alexander König
  - ...
-

# In a nutshell

## CLARIN in the CLASSROOM

### A Guide for Teachers and Students

#### 1. About the Authors

- 1.1 Context and Motivation
- 1.2 Aims of this Guide

#### 2. What are European Research Infrastructures?

#### 3. What is CLARIN?

#### 4. Accessing CLARIN

#### 5. How to Use CLARIN for Language and Linguistic Research

- 5.1 Searching and Finding Published Language Resources
- 5.2 Searching across Text Collections
- 5.3 Collecting and Citing Language Resources
- 5.4 Finding and Querying Large Collections of Corpora
- 5.5 Finding a Language Processing Tool or Service
- 5.6 Using Published Language Resources and Datasets
- 5.7 Archiving and Sharing Language Resources
- 5.8 Guidance on the Use of Standards and File Formats
- 5.9 Guidance on Licenses and Legal issues in Data Reuse

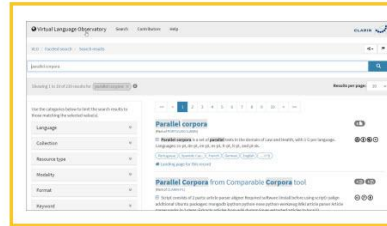
#### 6. How to Use the Knowledge Infrastructure

#### 7. Lesson Plans

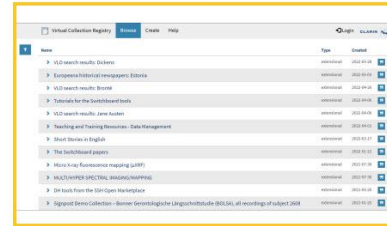
#### 8. Conclusions

#### Contribution and Maintenance

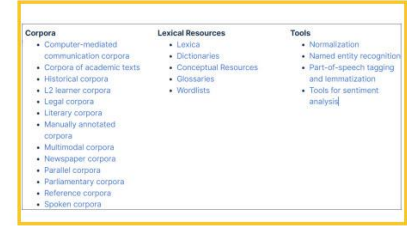
#### Bibliography



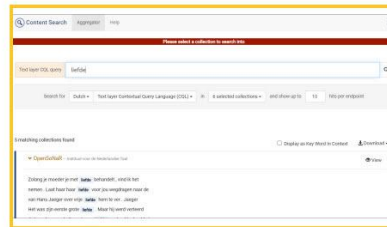
1 Find and(re)use published language resources



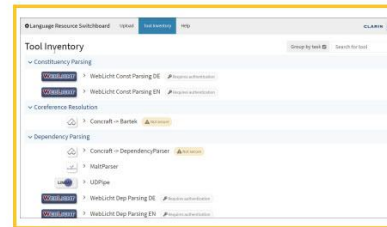
2 Collect, cite and share collections of virtual resources



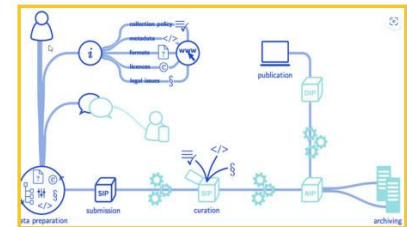
3 Find and query high-quality corpora



4 Search for specific patterns collections of resources



5 Find a matching tool to process your text file (s)



6 Archive and share your language resources

# Finding Language Resources

## CLARIN Virtual Language Observatory

- A catalogue that harvests **metadata** about language resources available in **distributed repositories**
- It does not contain language resources, just helps you locate it via **persistent identifiers**
- Even if a resource has a restricted access, the metadata is always freely accessible
- It uses **faceted search** to narrow down your searches
- **Useful as a first step in a project to identify whether, e.g., there is already a corpus published on the topic that the student has in mind**

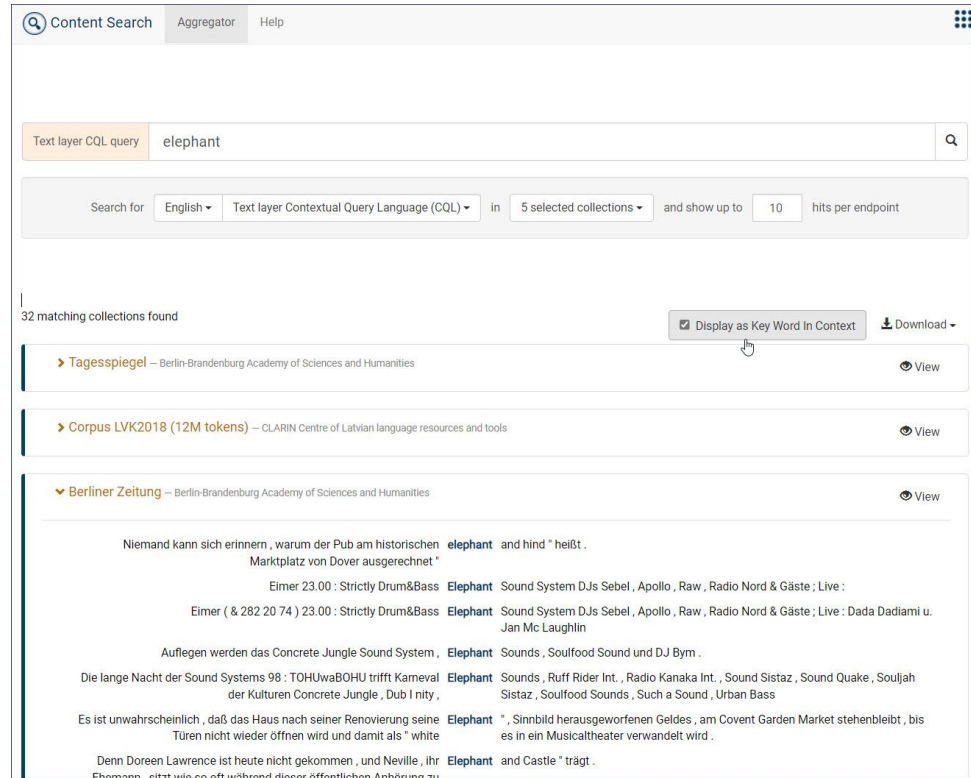
The screenshot shows the Virtual Language Observatory (VLO) search results page. The search criteria are 'tmx corpora' and 'Italian'. The page displays four search results, each with a brief description and a list of languages. The results are:

- Parallel Global Voices.** (Part of CLARINEL Catalogue) - Description: Parallel Global Voices is a set of parallel and monolingual corpora generated from the Global Voices multilingual group of websites (https://globalvoices.org/), where volunteers publish and translate news stories in more than 40 languages. The original content from the Global Voices websites is available by the authors... Languages: Arabic, Aymara, Bengali; Ban., Bulgarian, ... (+35)
- ECDC Translation Memory** (Part of CLARINEL Catalogue) - Description: In October 2012, the European Union (EU) agency 'European Centre for Disease Prevention and Control' (ECDC) released a translation memory (TM), i.e. a collection of sentences and their professionally produced translations, in twenty-five languages. The data gets distributed via the web pages of the EC's Joint Research ... Languages: Bulgarian, Czech, Danish, Spanish; Cas., Estonian, ... (+20)
- DGT-TM-2016** (Part of CLARINEL Catalogue) - Description: Since November 2007 the European Commission's Directorate-General for Translation has made its multilingual Translation Memory for the Acquis Communautaire, DGT-TM, publicly accessible in order to foster the European Commission's general effort to support multilingualism, language diversity and the re-use of Commission... Languages: Bulgarian, Czech, Danish, Spanish; Cas., Estonian, ... (+19)
- DGT-Translation Memory** (Part of CLARINEL Catalogue) - Description: Since November 2007 the European Commission's Directorate-General for Translation has made its multilingual Translation Memory for the Acquis Communautaire, DGT-TM, publicly accessible in order to foster the European Commission's general effort to support multilingualism, language diversity and the re-use of Commission...

# Finding Language Resources

## Content Search

- A search engine that helps the users locate specific linguistic patterns across several text collections
- The data itself stays at the centre where it is hosted
- It is not possible to rank the search results, but they can be downloaded in a variety of formats and perform further analysis in other tools
- **Useful as a first step to discover where interesting LRs are hosted**



The screenshot displays the 'Content Search' interface. At the top, there are navigation tabs for 'Content Search', 'Aggregator', and 'Help'. Below this is a search bar with the text 'elephant' and a search icon. Underneath the search bar, there are filters: 'Search for' set to 'English', 'Text layer Contextual Query Language (CQL)' selected, 'in 5 selected collections', and 'and show up to 10 hits per endpoint'. The results section shows '32 matching collections found'. A checkbox labeled 'Display as Key Word In Context' is checked. The results are listed as follows:


- Tagesspiegel** – Berlin-Brandenburg Academy of Sciences and Humanities (View)
- Corpus LVK2018 (12M tokens)** – CLARIN Centre of Latvian language resources and tools (View)
- Berliner Zeitung** – Berlin-Brandenburg Academy of Sciences and Humanities (View)

Below the collection headers, the search results are displayed in a list format, showing the word 'elephant' in context within various text snippets. For example, one snippet from 'Berliner Zeitung' reads: 'Niemand kann sich erinnern, warum der Pub am historischen Marktplatz von Dover ausgerechnet \* elephant and hind \* heißt.'

# Finding Language Resources

## Discipline-specific repositories

- Preserve, manage, and provide access to language resources in a variety of formats
- Continuous curation
- Language resources described with consistent metadata
- Data can be cited
- E.g. [CLARIN B-Centres](#), [The Open Language Archive](#), [Meta-Share](#)


Repository details 

### ILC-CNR for CLARIN-IT repository

General **Institutions** Terms Standards

Name of repository	ILC-CNR for CLARIN-IT repository
Additional name(s)	ILC4CLARIN CLARIN-IT Repository
Repository URL	<a href="https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/">https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/</a>
Subject(s)	<a href="#">Linguistics</a> <a href="#">Humanities</a> <a href="#">Humanities and Social Sciences</a> <a href="#">Artificial Intelligence, Image and Language Processing</a> <a href="#">Computer Science</a> <a href="#">Computer Science, Electrical and System Engineering</a> <a href="#">Engineering Sciences</a>
Description	ILC-CNR for CLARIN-IT repository is a library for linguistic data and tools. Including: Text Processing and Computational Philology; Natural Language Processing and Knowledge Extraction; Resources, Standards and Infrastructures; Computational Models of Language Usage. The studies carried out within each area are highly interdisciplinary and involve different professional skills and expertises that extend across the disciplines of Linguistics, Computational Linguistics, Computer Science and Bio-Engineering.
Contact	dspace-clarin-it-ilc-help@ilc.cnr.it Alessandro Enea@ilc.cnr.it
Content type(s)	<a href="#">Standard office documents</a> <a href="#">Plain text</a> <a href="#">Databases</a> <a href="#">Software applications</a> <a href="#">Scientific and statistical data formats</a>
Certificates and Standards	CoreTrustSeal CLARIN certificate B
Keyword(s)	<a href="#">Ancient Greek</a> <a href="#">Italian Language</a> <a href="#">Bioengineering</a> <a href="#">computational linguistics</a> <a href="#">computational philology</a> <a href="#">corpora</a> <a href="#">treebanks</a>
Repository type(s)	disciplinary
Mission statement for designated community	<a href="https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/about?locale-attribute=en">https://dspace-clarin-it.ilc.cnr.it/repository/xmlui/page/about?locale-attribute=en</a>
Research data repository language(s)	English
Data and/or service provider	data provider

[Back to search](#) [Submit a change request](#) [Get a badge](#)

 Cite this re3data.org record:  
re3data.org: ILC-CNR for CLARIN-IT repository, editing status 2022-01-03, re3data.org - Registry of Research Data Repositories. <http://doi.org/10.17616/R3W365> last accessed: 2022-06-30

<http://doi.org/10.17616/R3W365>

# Find and Query Large Collections of Corpora

## CLARIN Resource Families

- User-friendly overviews of well-curated corpora and tools
- Download or query with concordancers, e.g. [Korp](#), [Corpuscle](#) and [KonText](#)

Corpora	Lexical Resources	Tools
<ul style="list-style-type: none"><li>• Computer-mediated communication corpora</li><li>• Corpora of academic texts</li><li>• Historical corpora</li><li>• L2 learner corpora</li><li>• Literary corpora</li><li>• Manually annotated corpora</li><li>• Multimodal corpora</li><li>• Newspaper corpora</li><li>• Parallel corpora</li><li>• Parliamentary corpora</li><li>• Reference corpora</li><li>• Spoken corpora</li></ul>	<ul style="list-style-type: none"><li>• Lexica</li><li>• Dictionaries</li><li>• Conceptual Resources</li><li>• Glossaries</li><li>• Wordlists</li></ul>	<ul style="list-style-type: none"><li>• Normalization</li><li>• Named entity recognition</li><li>• Part-of-speech tagging and lemmatization</li><li>• Tools for sentiment analysis</li></ul>

# Find and Query Large Collections of Corpora

Croatian Twitter training corpus  
ReLDI-NormTagNER-hr 2.0

**Size:** 89,000 tokens

**Annotation:** tokenisation, sentence segmentation, word normalisation, morphosyntactic tagging, lemmatisation and Named Entity recognition

**Licence:** CC BY 4.0

Croatian

This corpus contains Tweets. The corpus is morphosyntactically tagged following the MULTEXT-East Version 4 tagset.

The corpus is available through the concordancers KonText and noSketchEngine and for download from the CLARIN.SI repository.

For the relevant publication, see [Miličević and Ljubešić \(2016\)](#)

KonText

noSketch

Download

[Query the ReLDI corpus via NoSketch engine on CLARIN.SI](#)

The screenshot shows the NoSketch Engine interface. At the top, there is a search bar with a dropdown menu set to 'ReLDI-hr (manually tagged Croatian tweets)'. Below the search bar, the main content area displays the corpus details for 'ReLDI-hr (manually tagged Croatian tweets)'. The interface is divided into several sections: 'Counts', 'General info', 'Lexicon sizes', and 'Tags legend'. The 'Counts' section shows 89,104 Tokens, 71,768 Words, 7,939 Sentences, and 3,871 Documents. The 'General info' section shows the Corpus description (Document), Language (Croatian), Encoding (UTF-8), Compiled date (09/11/2019 15:41:23), and Tagset (Description). The 'Lexicon sizes' section shows word counts for various tags: word (27,289), norm (25,395), lemos (17,270), tag (694), ud\_pos (90), ud\_feats (764), diff (5), lc (25,219), lemma (16,675), and lemma\_lc (16,020). The 'Tags legend' section shows a list of tags and their corresponding parts of speech: Noun (N.\*), Noun proper (Np.\*), Noun common (Nc.\*), Verb (V.\*), Adjective (A.\*), Pronoun (P.\*), Adverb (R.\*), Preposition (S.\*), and Conjunction (C.\*). At the bottom, the 'Structures and attributes' section shows 'text 3,871' and 'name 5,883'.

NoSketch Engine

ReLDI-hr (manually tagged Croatian tweets)

## ReLDI-hr (manually tagged Croatian tweets)

Croatian tweets with manually normalised (standardised), morphosyntactically tagged and lemmatised words and named entities ReLDI-hr v2.1

Counts	General info	Lexicon sizes	Tags legend
Tokens 89,104	Corpus description <a href="#">Document</a>	word 27,289	Noun N.*
Words 71,768	Language Croatian	norm 25,395	Noun proper Np.*
Sentences 7,939	Encoding UTF-8	lemos 17,270	Noun common Nc.*
Documents 3,871	Compiled 09/11/2019 15:41:23	tag 694	Verb V.*
	Tagset <a href="#">Description</a>	ud_pos 90	Adjective A.*
		ud_feats 764	Pronoun P.*
		diff 5	Adverb R.*
		lc 25,219	Preposition S.*
		lemma 16,675	Conjunction C.*
		lemma_lc 16,020	

### Structures and attributes

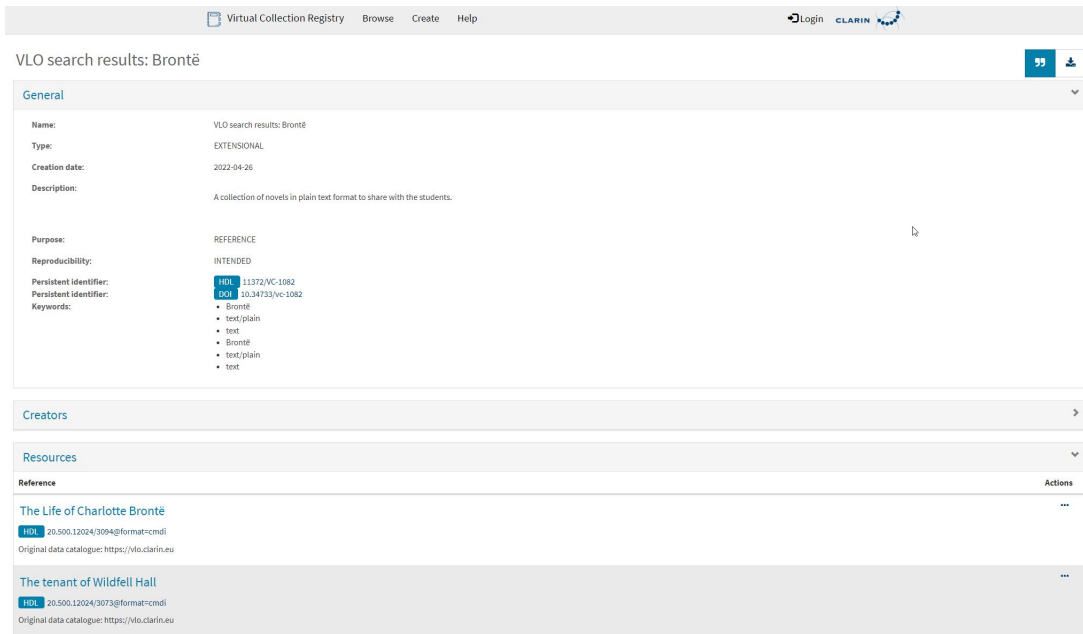
text 3,871	⌵
name 5,883	⌵



# Collecting and Citing Language Resources

## Virtual Collection Registry

- Collect the datasets discovered in the VLO or other data catalogues into a Virtual Collection
- Share the collection with the others
- Cite the collection
- Process the collection with Switchboard tools



The screenshot displays the Virtual Collection Registry (VLO) search results for the query 'Brontë'. The interface includes a navigation bar at the top with 'Virtual Collection Registry', 'Browse', 'Create', and 'Help' menus, along with a 'Login' button and the CLARIN logo. The search results are organized into several sections:

- VLO search results: Brontë**: A summary section with a '99' badge and a user icon.
- General**: A detailed view of the collection with the following metadata:
  - Name: VLO search results: Brontë
  - Type: EXTENSIONAL
  - Creation date: 2022-04-26
  - Description: A collection of novels in plain text format to share with the students.
  - Purpose: REFERENCE
  - Reproducibility: INTENDED
  - Persistent Identifier: [DOI 10.34733/vc-1082](https://doi.org/10.34733/vc-1082)
  - Keywords: Brontë, text/plain, text, Brontë, text/plain, text
- Creators**: A section with a right-pointing arrow.
- Resources**: A section with a dropdown arrow and an 'Actions' column. It lists two resources:
  - The Life of Charlotte Brontë**: [DOI 10.500.12024/3094#format=cmdi](https://doi.org/10.500.12024/3094#format=cmdi), Original data catalogue: <https://vlo.clarin.eu>
  - The tenant of Wildfell Hall**: [DOI 10.500.12024/3073#format=cmdi](https://doi.org/10.500.12024/3073#format=cmdi), Original data catalogue: <https://vlo.clarin.eu>

# Processing Text Collections

## The Language Resource Switchboard

- A service that allows you to find a matching tool to analyse plain text files
- Taggers, lemmatizers, named entity recognizers, chunking tools etc.

The screenshot displays the Virtual Language Observatory (VLO) interface. The main content area shows search results for the record 'Jane Eyre / by Charlotte Bronte [sic]'. A table lists the following files:

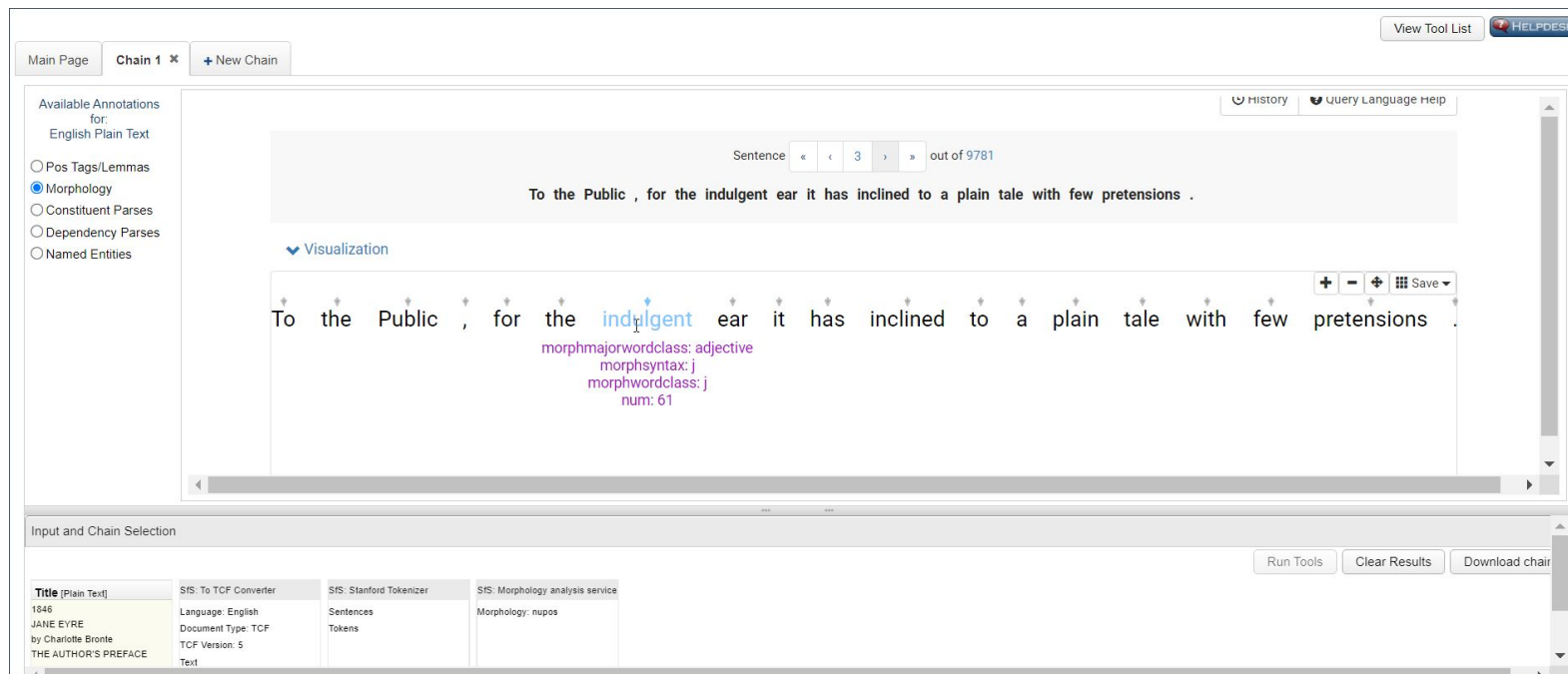
Name	Type
HDL_2001	Landing page
dublin_core.xml	XML
metadata_local.xml	XML
header2001.xml	XML
jeyre_2001.bt	Plain Text

An overlay window titled 'Switchboard' is open, showing details for the file 'jeyre\_2001.bt' (1007.02 KiB). It includes fields for 'Mediatype' (text/plain) and 'Language' (English). Under the 'Matching Tools' section, there is a search bar and a list of tools:

- Constituency Parsing
  - WebLight Const Parsing EN (Open, Requires authentication)
- Dependency Parsing
  - UDPipe (Open)

# Processing Text Collections

Example of morphological analysis in [WebLicht](#)



The screenshot displays the WebLicht interface for morphological analysis. The main window shows the sentence "To the Public, for the indulgent ear it has inclined to a plain tale with few pretensions." with the word "indulgent" highlighted in blue. Below the word, the following morphological information is displayed:

- morphmajorwordclass: adjective
- morphsyntax: j
- morphwordclass: j
- num: 61

The interface includes a sidebar with "Available Annotations for English Plain Text" and options for "Pos Tags/Lemmas", "Morphology" (selected), "Constituent Parses", "Dependency Parses", and "Named Entities". The bottom section, "Input and Chain Selection", shows a table of processing steps:

Title [Plain Text]	SIS: To TCF Converter	SIS: Stanford Tokenizer	SIS: Morphology analysis service
1846 JANE EYRE by Charlotte Brontë THE AUTHOR'S PREFACE	Language: English Document Type: TCF TCF Version: 5 Text	Sentences Tokens	Morphology: nupos

Buttons for "Run Tools", "Clear Results", and "Download chain" are visible at the bottom right.

# Archiving Language Resources

- Institutional repository
- Domain-specific repository
- Infrastructural repository: [Depositing Services | CLARIN ERIC](#)
- Support with data preparation, license selection, depositing process

## Item submission



Example of [data submission lifecycle in CLARIN.SI](#)

# Sharing Language Resources

- Select appropriate licenses
- The repository assigns unique identifiers
- Others can cite the resources

Automatically stress labelled morphological lexicon Sloleks 1.2, version 1.1 

 Please use the following text to cite this item or export to a predefined format:

BIBTEX

CMDI

Krsnik, Luka; Robnik-Šikonja, Marko; Šef, Tomaž and Krek, Simon, 2018, *Automatically stress labelled morphological lexicon Sloleks 1.2, version 1.1*, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, <http://hdl.handle.net/11356/1186>.



Share:  

# Guidance on the Use of Standards and Formats

## CLARIN Standards Information System

- Data deposition formats
- Language-technology related standards



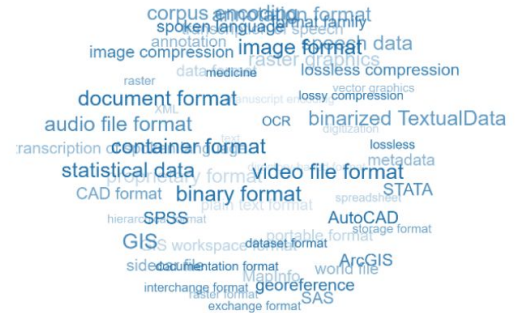
Home  
Centres  
Format Recommendations  
Data Deposition Formats  
Functional Domains  
File Extensions  
Media Types  
Statistics  
Popular Formats  
Sanity Check  
Standards and Specifications  
Standard Bodies  
Topics  
Search  
API  
Contact  
  
Login  
Register

> Home

### CLARIN Standards Information System

The primary role of the CLARIN Standards Information System is, currently, to aggregate and visualize the list of recommendations for data deposition formats, specified by CLARIN centres that offer deposition services (mostly the so-called B-centres). That list is available in the "Format Recommendations" section, and its various logical subcomponents can be accessed from the menu on the left.

The keyword cloud below provides an alternative way to access the format and recommendation information. It is still in the process of being fine-tuned.



# Guidance on Licenses and Legal Issues in Data Reuse

CLARIN Legal Information Platform:

- [Introduction to Copyright and Related Rights](#)
- [Licensing Practice](#)
- [Personal Data Protection](#)

# Would you like to learn more?

We can share a draft of the guide with you

Test the services and see how you can use them in the classroom

Contribute to the guide with examples of learning activities

**Email:** [iulianna@clarin.eu](mailto:iulianna@clarin.eu)



**14:40 - 12:30**

**Demonstrations of the  
UPSKILLS Learning Content  
Blocks**

Introduction to Language Data:  
Standards and Repositories

---

# 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

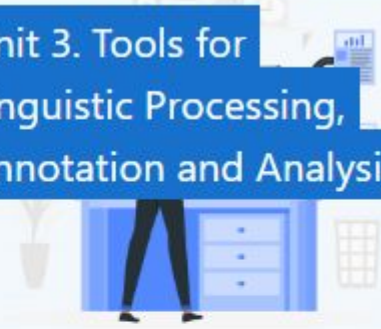
Unit 1. Introduction to  
Research Infrastructures  
of Language Resources  
and ...



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 Findability, Accessibility, Interoperability, and Reuse 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](https://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.

