CLARIN ERIC, a distributed infrastructure for language resources and technologies and its national node CLARIN-IT

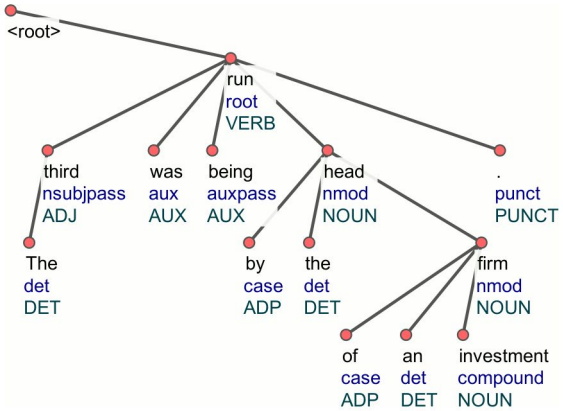
Francesca Frontini - *CLARIN Board of Directors*
Monica Monachini - CLARIN-IT National Coordinator
Istituto di Linguistica Computazionale - ILC CNR
Venice, 22/09/2022

Materials for this lesson:
https://docs.google.com/document/d/1sNQ0eIwJFFAInGm00t7Lb2tytOorzjzoOs796CR8fHQ/edit?usp=sharing

CLARIN & Language Resources

CLARIN

Common Language Resources and Technology Infrastructure



https://www.clarin.eu/

1. Intro
2. The CLARIN Research Infrastructure
    Exercise 1 - Finding and processing data
3. Knowledge Sharing
    Exercise 2 - The ParlaMint corpora
4. The CLARIN-IT national consortium
    Exercise 3 - Explore ILC4CLARIN
5. A Cluster of SSH research Infrastructures
    Exercise 4 - The SSHOC Marketplace

Intro: Quality Language Resources for NLG

Language Resources

The term language resources refers to a broad range of **speech and language data types in machine readable form**, as well as tools and services for the processing of language data. Following a longstanding tradition (Godfrey & Zampolli 1997), the term language resources **also covers software tools for the preparation, collection, management, or use of other resources**. Examples of such tools are corpus management and exploration systems, OCR systems, pipelines, speech processing systems, machine translation systems, environments for manual annotation and evaluation.

Building blocks for NLP/NLG

Language resources are used in a number of downstream tasks including NLG related ones:

- Reference corpora
- Language models
- NLP annotation pipelines
- Morphological analysers
- Machine translation tools
- Corpus exploration tools

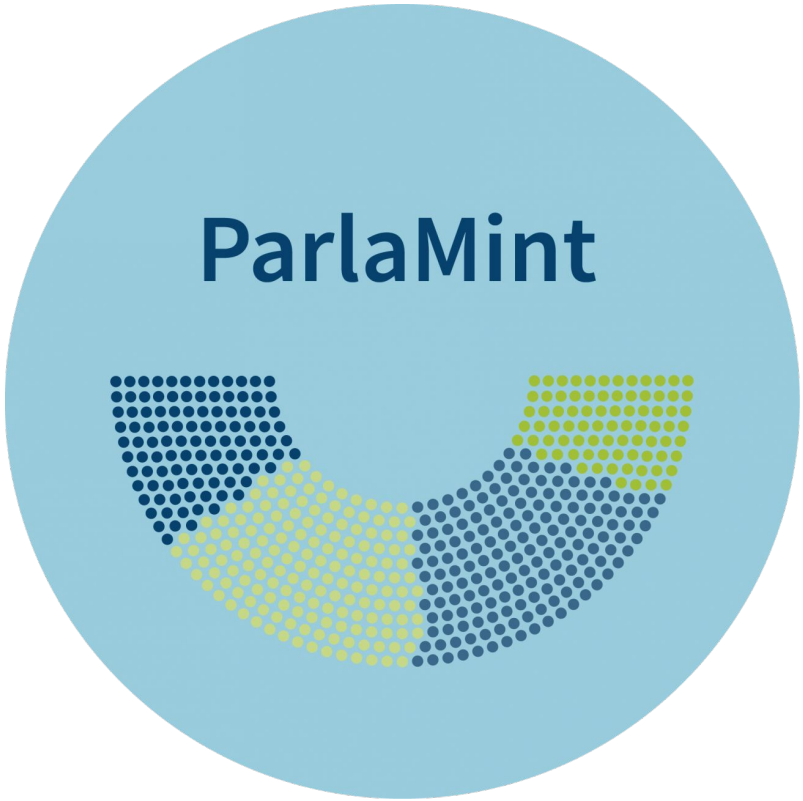
Availability of such LRs is crucial. These should be available in **different languages**, including less-represented ones.

ParlaMINT

Comparable parliamentary corpora

17 European parliaments with almost half a billion words

The complete corpora available in CLARIN in TEI **hdl.handle.net/11356/1432** and with added linguistic annotations: **hdl.handle.net/11356/1431**



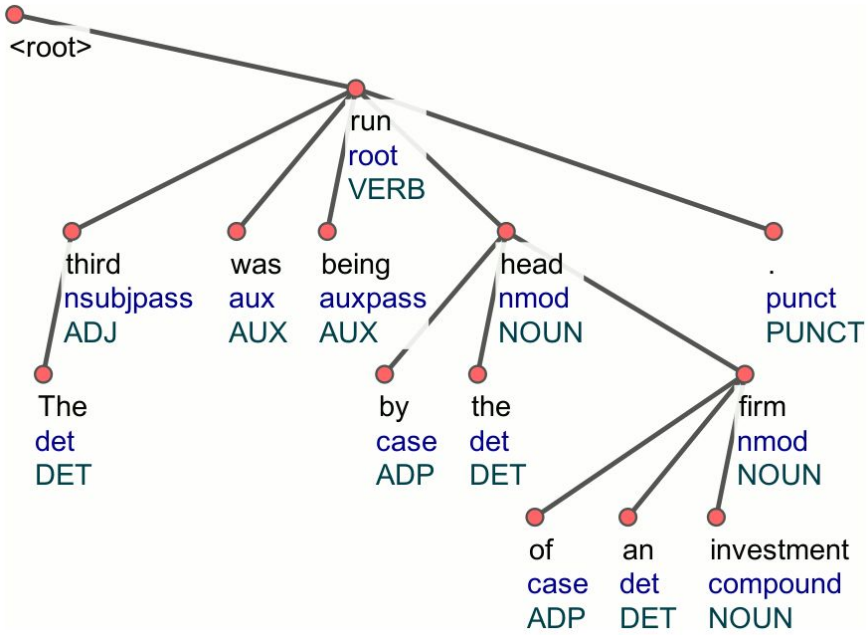
Source https://www.clarin.eu/parlamint

UDpipe

The **Universal Dependencies 2.10 models** contain 123 models of 69 languages, each consisting of a tokenizer, tagger, lemmatizer and dependency parser, all trained using the UD data.

The models are based on Universal Dependencies 2.10 treebanks.

Available as REST services via LINDAT CLARIAH
https://lindat.mff.cuni.cz/services/udpipe



Source https://ufal.mff.cuni.cz/udpipe

EMBEDDIA

Cross-Lingual Embeddings for Less-Represented Languages in European News Media

The EMBEDDIA project seeks to address these challenges by leveraging innovations in the use of **cross-lingual embeddings** coupled with deep neural networks to allow existing monolingual resources to be **used across languages**, leveraging their high speed of operation for near real-time applications, without the need for large computational resources.

Tools made available via CLARIN Slovenia

Source http://embeddia.eu/

Reference corpora

DK-CLARIN Reference Corpus of General Danish **Size:** 45.1 million words **Annotation:** PoS-tagged, sentence and paragraph segmentation, lemmatized **Licence:** CLARIN ACA-NC	Danish	This corpus includes Danish texts published between 2008 and 2011. The corpus is encoded in TEI. Non-linguistic metadata includes information on source and year of publication. The corpus is available for download from the CLARIN-DK repository.	Download

Source https://www.clarin.eu/resource-families/reference-corpora

FAIR Data

FAIR Principles

Compliance

Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.

- **F1**. Resource is uploaded to a public repository.
- **F2**. Metadata are assigned a globally unique and persistent identifier.

Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.

- **A1**. Resource is accessible for download or manipulation by humans and is ideally also machine readable.
- **A2**. Publications and data repositories have contingency plans to assure that metadata remain accessible, even when the resource or the repository are no longer available.

Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.

- **I1**. Resource is uploaded to a repository that is interoperable with other platforms.
- **I2**. Repository meta- data schema maps to or implements the CG Core metadata schema.
- **I3**. Metadata use standard vocabularies and/or ontologies.

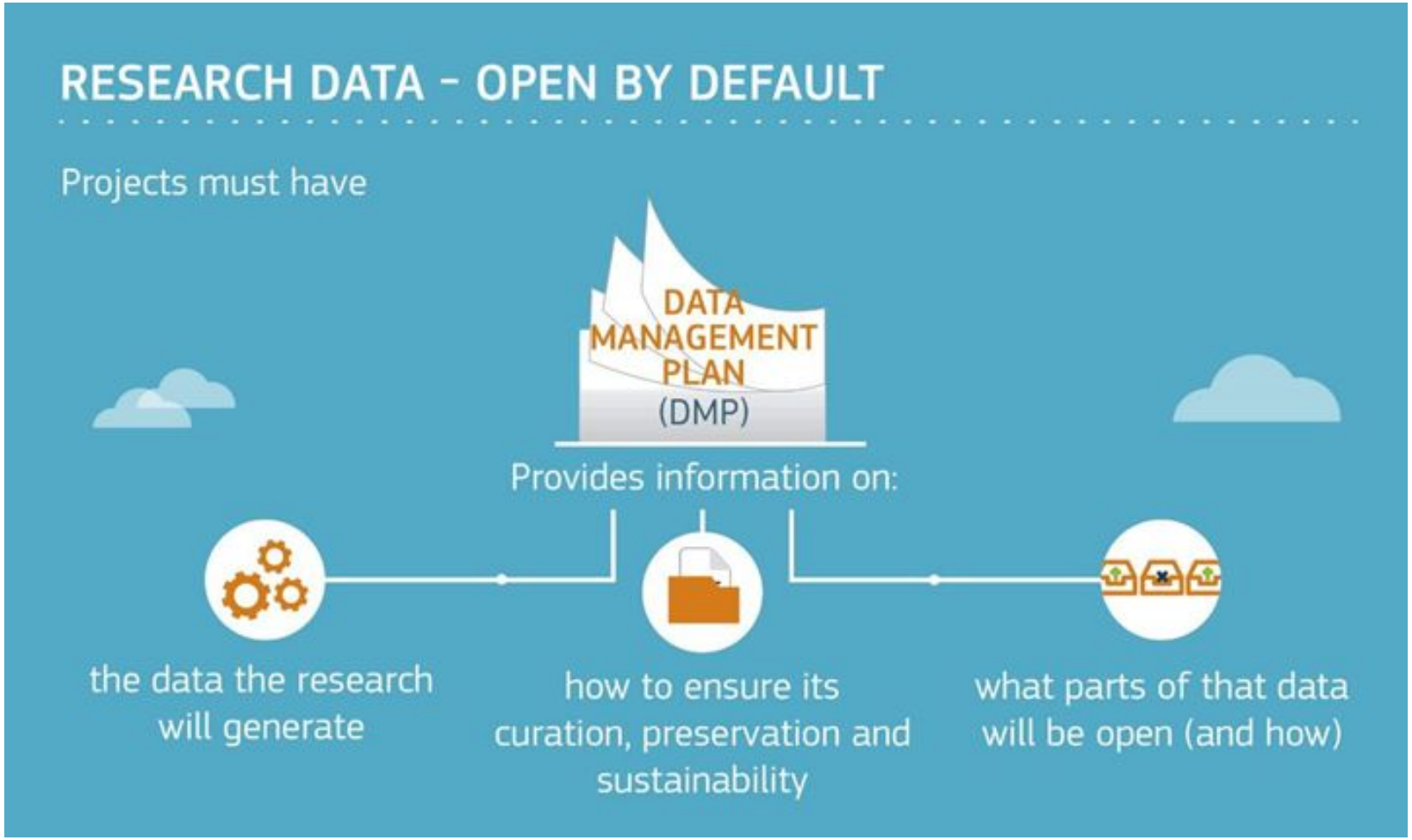
Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines and humans.

- **R1.** Metadata are released with a clear and accessible usage license.
- **R2.** Metadata about data and datasets are richly described with a plurality of accurate and relevant attributes.

'FAIR Guiding Principles for scientific data management and stewardship', 2016
https://www.go-fair.org/fair-principles/

A European Strategy

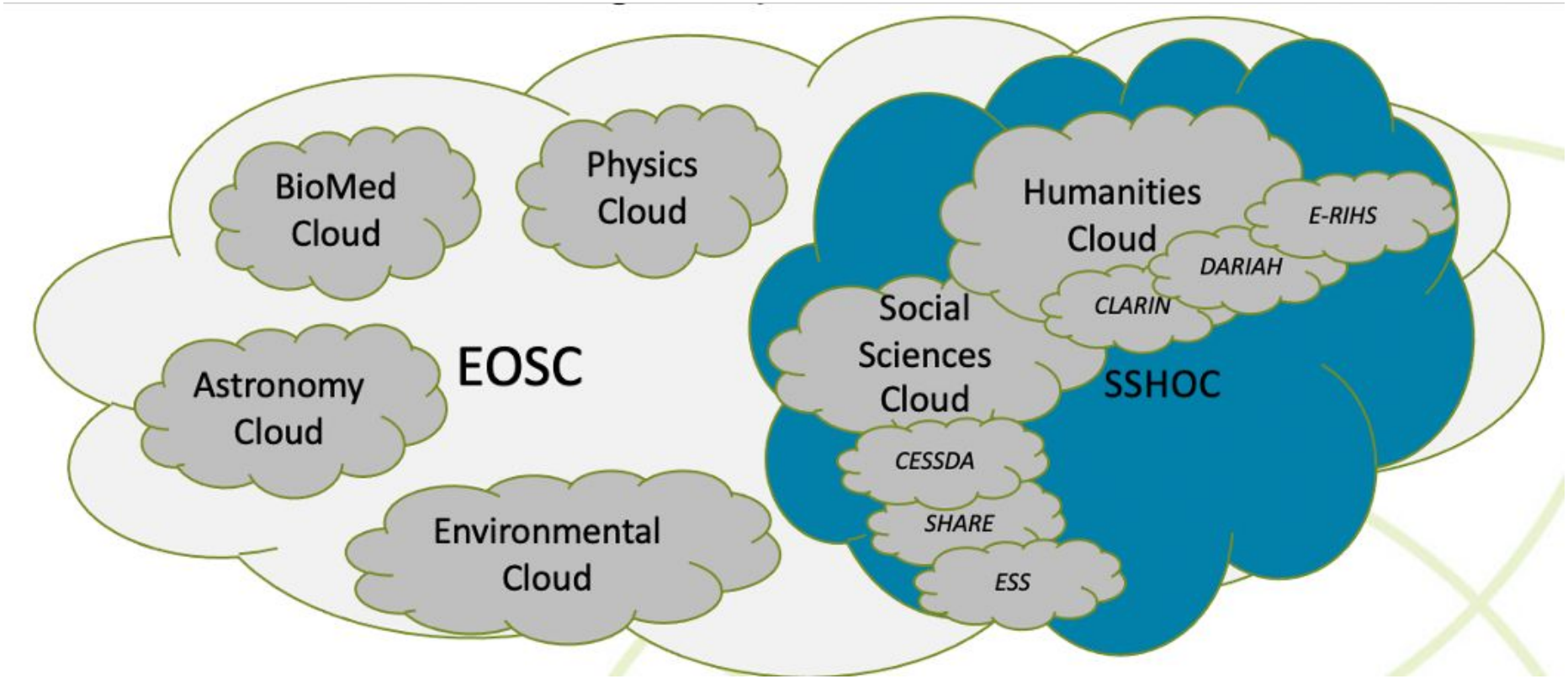


Source: https://www.openaire.eu/what-is-a-data-management-plan

How to ensure ...

- long term deposit
- metadata quality
- findability
- citability
- clear licenses
- interoperability?

An Ecosystem of Infrastructures



What you will learn today

- How to make best use of this ecosystem
  - find and deposit data
  - find and use tools
  - get support and exchange knowledge within CLARIN and the SSH Open Cluster

Research Infrastructures

European Commission defines research infrastructures as **facilities** that provide **resources and services for research communities to conduct research and foster innovation**. They can be used beyond research, e.g. for education or public services and they may be single-sited, distributed, or virtual.

SSH RIs

In the age of Artificial Intelligence and the so-called “big data society,” large Research Infrastructures (RIs) in the Arts and Humanities, such as CLARIN and DARIAH, provide scholars, researchers, students and educators with the **knowledge, high-quality data, services and technologies** to help them explore the global scientific and societal transformations, and tackle the new challenges.

The CLARIN Research Infrastructure for data sharing

CLARIN in a nutshell

- has the **ESFRI ERIC** status since 2012, **Landmark** since 2016
- provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to digital language data (in written, spoken, video or multimodal form)
  - and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
  - through a **single sign-on** environment
- serves as an ecosystem for **knowledge sharing and training**
- is an integral part of **the European Open Science Cloud**
  - See **clarin.eu/eosc**

CLARIN data and communities

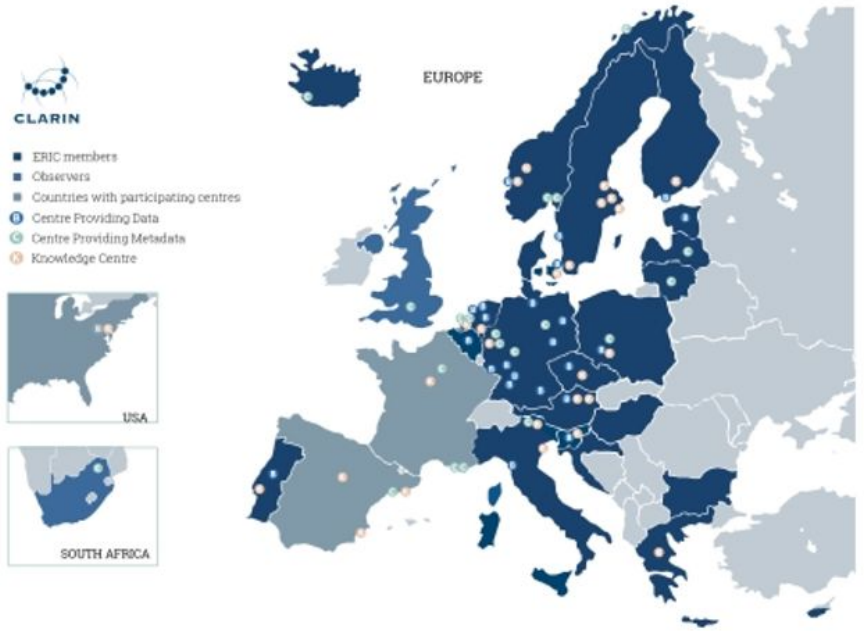
- Newspaper archives
- Literary texts
- Parliamentary records
- Literary texts
- Historical letters
- Broadcast archives
- Oral History data
- Social Media data
- L-2 Learner Resources
- Survey data
- ...

For the CLARIN Resource Families initiative, see: https://www.clarin.eu/resource-families

- Digital humanities
- Linguistics and Philology
- Translation and Lexicography
- Literary Studies
- History
- Political and Social Sciences
- Media Studies
- Culture, Folklore, Anthropology
- Speech therapy
- General Public
- ...

CLARIN today

- **21 members**: (AT, BE, BG, CY, CZ, DE, DK, EE, FI, GR, HR, HU,IS, IT, LT, LV, NL, NO, PL, PT, SE, SI)
- 2 observers: UK, ZA
- > **70 centres**



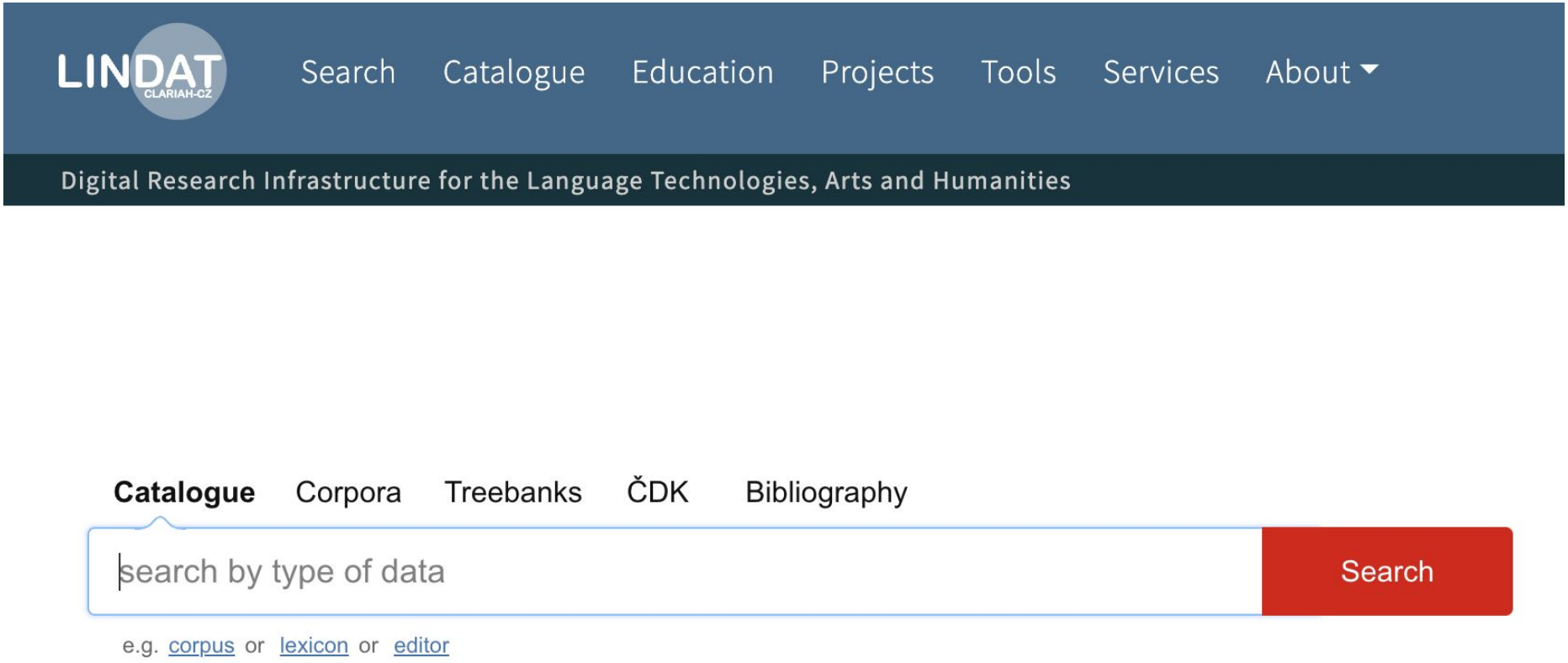
CLARIN
Common Language Resources and Technology Infrastructure

CLARIN types of centres

- **B-Centres**
  - These are the distributed data centres that universities or academic institutions can access to search for or deposit language resources, access services and expert knowledge. There are several certified B-centres that have passed the CLARIN centre assessment procedure. The CLARIN-B centres are required to apply for the CoreTrustSeal certification.
- **C-Centres**
  - These are centres that provide metadata to CLARIN but they do not offer any other services.
- **K-Centres**
  - Knowledge centres that share knowledge and expertise on one or more aspects of the domains covered by the CLARIN infrastructure.

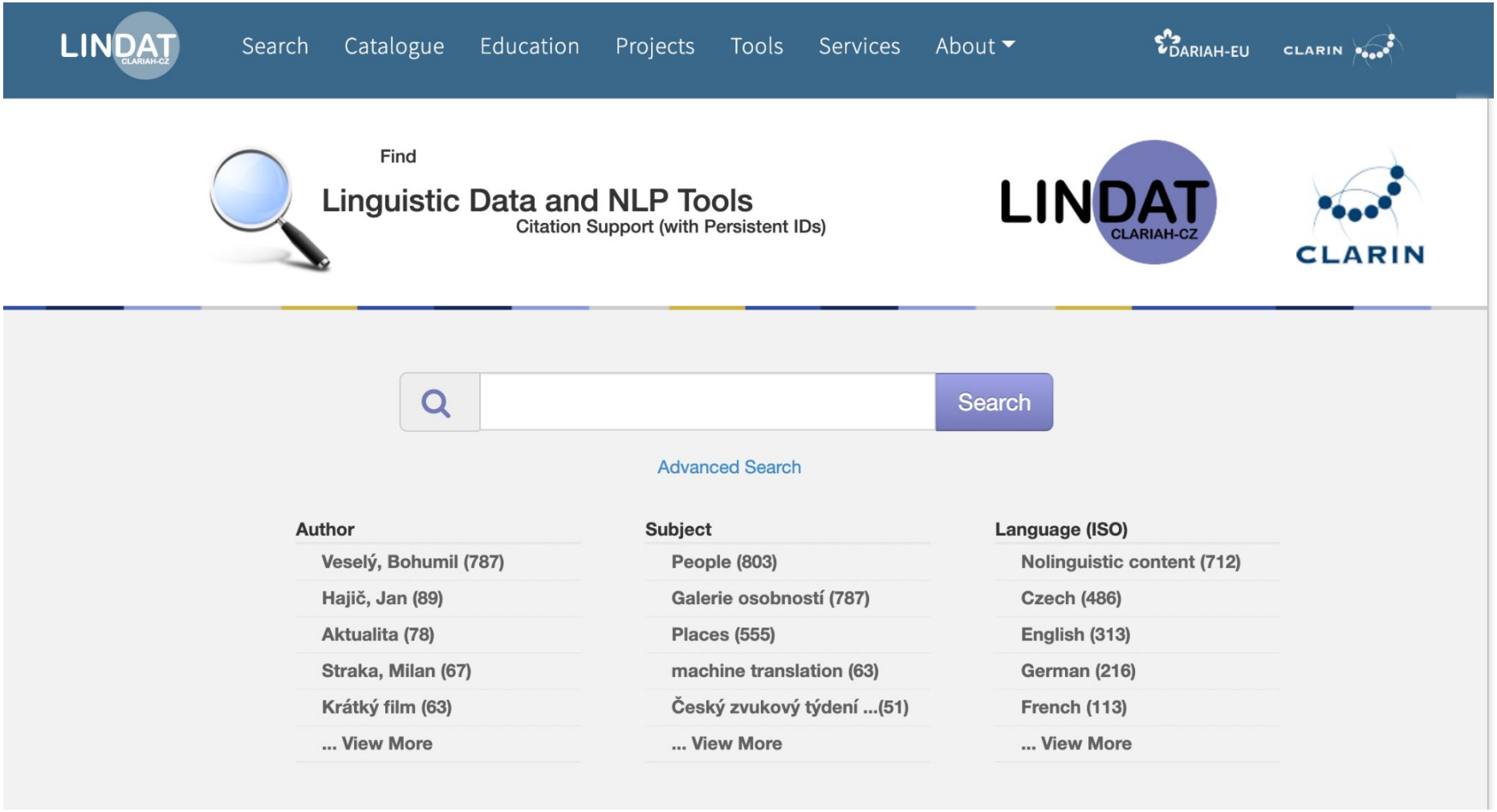
https://www.clarin.eu/content/overview-clarin-centres

....Consortia and Centres offer Data, Tools, Services



https://lindat.cz/

....hosting repositories



https://lindat.mff.cuni.cz/repository/xmlui/?locale-attribute=en

.... offering tools for exploration and processing

Tools for Natural Language Processing

Terms of use
Service Status

Enrich your texts
Make them more searchable

Phonetics, Phonology

– UWebASR (audio transcription, Czech)

Machine Translation

– LINDAT translation

Morphology and tagging

– ElixirFM (Arabic)
– MorphoDiTa (Czech)
– UDPipe

Natural language processing

– Treex (Czech, English, Latin)
– UDPipe

Syntactic parsing

– Korektor (spellchecker, Czech)
– Parsito
– UDPipe

Search

– CzEngVallex (lexicon, Czech, English)
– EngVallex (lexicon, English)
– PDT-Vallex (lexicon, Czech)
– Dialogy.org (audio-visual corpora search)
– Internet Language Reference Book (Czech)
– KonText (concordances, collocations, word frequencies)
– TEITOK (search, visualize, edit corpora)
– PML Tree Query search (treebank search)

Discourse, Pragmatics

– Evald (coherence evaluation, native speakers of Czech)
– Evald (coherence evaluation, non-native speakers of Czech)
– KER (keyword extraction, Czech, English)
– NameTag (named-entity recognition, Czech, English)

https://lindat.cz/#tools

CLARIN-UK

Infrastructure for Digital Language Resources and Tools

Home | Centres | Resources | People | News | Events | Courses | Blog | About

Read more about CLARIN-UK

Data & Tools

More

CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh

Latest News

More

New members of the CLARIN_UK consortium

18 November 2020
Welcome aboard!

Events

More

Lancaster Symposium on Innovation in Corpus Linguistics 2021

Wednesday 23 June, Online

Corpus

https://www.clarin.ac.uk/

CLARIN-UK

Infrastructure for Digital Language Resources and Tools

Home | Centres | Resources | People | News | Events | Courses | Blog | About

- #LancsBox
- BNC
- CKLD
- CLAWS
- CLiC
- CorCenCC
- CQPweb
- ELAR
- GATE
- GATE Cloud
- Hansard at Huddersfield

Read more about CLARIN-UK

Data & Tools

More

CorCenCC

Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh

More

Events

More

Lancaster Symposium on Innovation in Corpus Linguistics 2021

Wednesday 23 June, Online

Centre for Corpus Research

Corpus Linguistics

https://www.clarin.ac.uk/

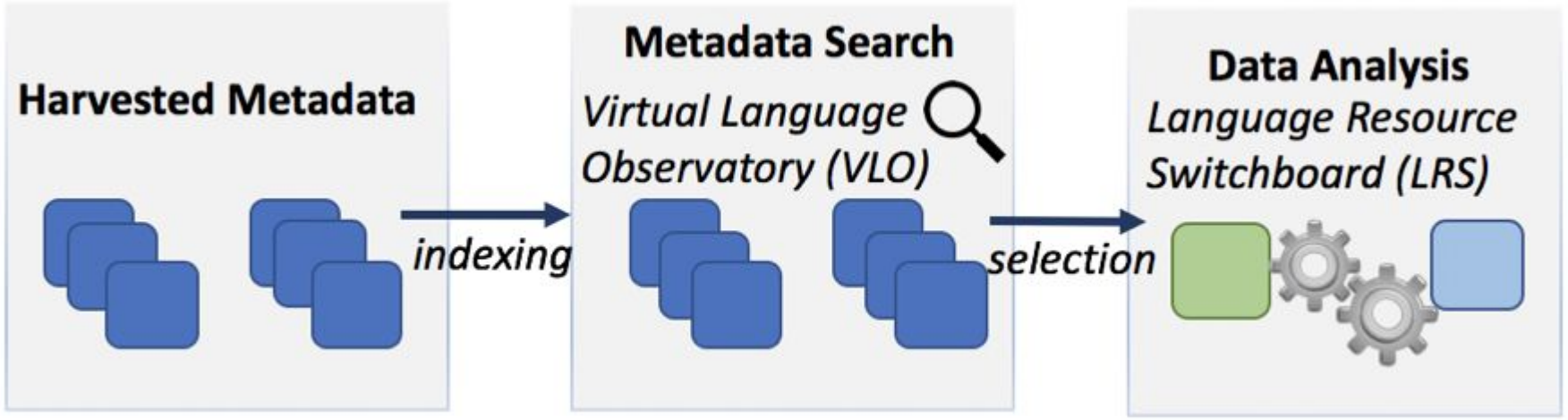
CLARIN Central Services

The CLARIN portal:
A single access point to the distributed infrastructure

- Make it easier for researchers, lecturers, students to **find** data, tools, training materials from centres
- Make it easier to **share and publish** data following open science practices
- Allow you to be **cited** for your work and acknowledge that of others
- Promote **knowledge sharing** and networking

https://www.clarin.eu/

The technical infrastructure



Findable Accessible Interoperable Reusable

clarin.eu/fair

vlo.clarin.eu

switchboard.clarin.eu

Barack Obama's identity-building in the health care debate: A corpus-assisted discourse study

AUTHOR
Katherina Riesner

Summary, in English

In this study, I demonstrate that identity-building is an important discursive strategy for President Barack Obama in the seven-year long debate surrounding the Affordable Care Act (ACA). The data for the study comes from a 6-million word corpus of speeches that were held by Obama between January 2009 and January 2016, all published by the White House. The speeches are classified according to genre, audience, topic and date of delivery. Throughout the paper, I adopt the notion that identity is intentionally constructed by the speaker and strategically exploited for his communicative goals. With the help of two methodological approaches, I investigate what kind of identities Obama builds. The purely qualitative part of the study deals with three central corpus speeches from a discourse-analytic perspective. In the second, more quantitative part, I use a group of seven verbs with epistemic meaning to trace the usage of two predominant discursive identities in the ACA debate. The results suggest that President Obama repeatedly constructs the identities of father and teacher to persuade his audience. I argue that his use of these identities constitutes an attempt to reach the argumentative goals of effectiveness and reasonableness.

Department/s
Master's Programme: Language and Linguistics

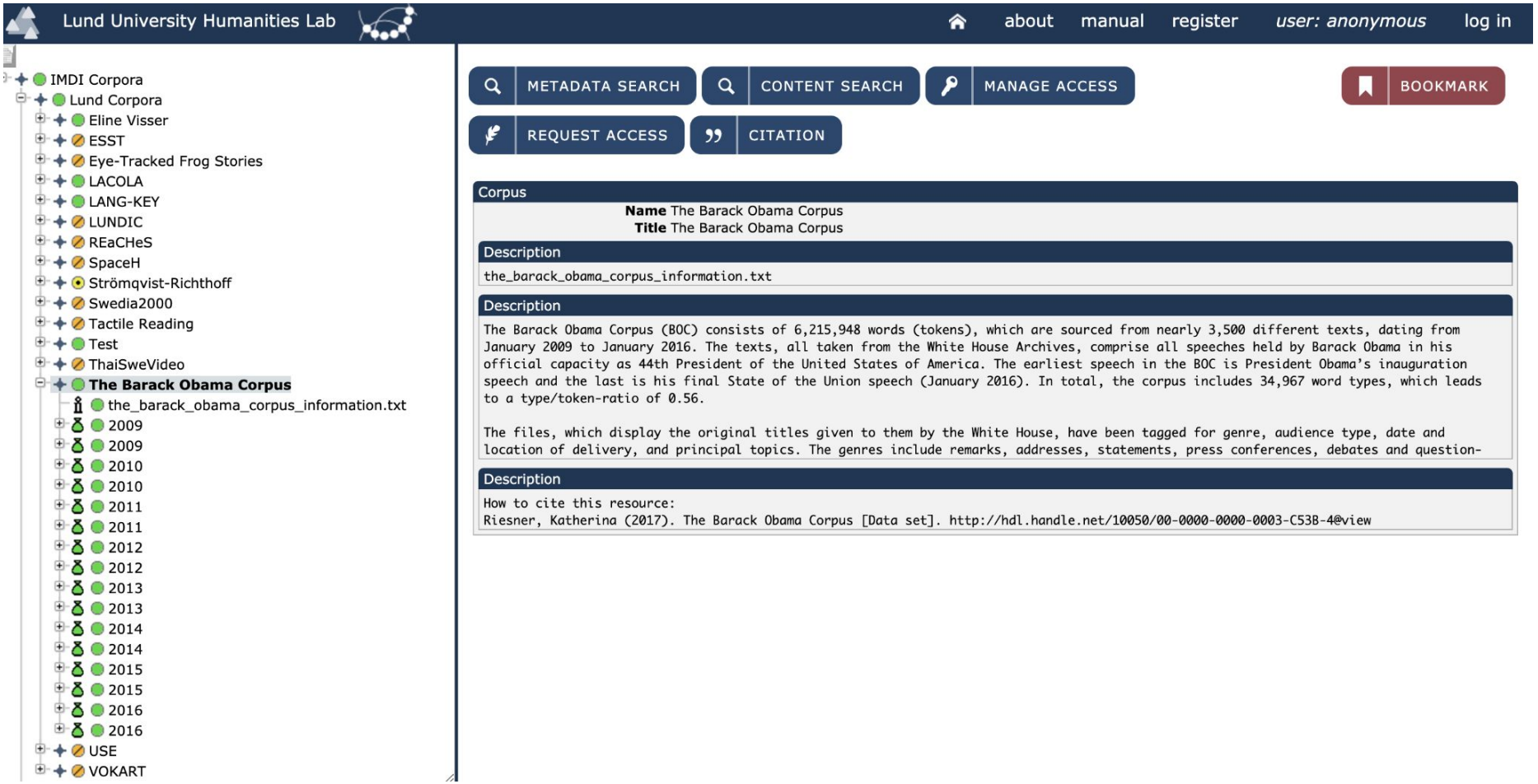
Publishing year
2016

Language
English

Full text
Available as PDF - 2 MB
Download statistics

Document type
Student publication for Master's degree (two years)

Topic
Languages and Literatures



Appropriate data citation and PID

Riesner, Katherina (2017). The Barack Obama Corpus [Data set].
http://hdl.handle.net/10050/00-0000-0000-0003-C53B-4@view

Virtual Language Observatory Search Contributors Help CLARIN

VLO / Faceted search / Search results

obama

Showing 5 results for obama

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Modality

The Barack Obama Corpus

(Part of Lund University Humanities Lab)

the_barack_obama_corpus_information.txt; The Barack **Obama** Corpus (BOC) consists of 6,215,948 words (tokens), which are sourced from nearly 3,500 different texts, dating from January 2009 to January 2016. The texts, all taken from the White House Archives, comprise all speeches held by Barack **Obama** in his official capac…

Landing page for this record at *corpora.humlab.lu.se*

SWE-CLARIN

metadata

vlo.clarin.eu

VLO and NLG resources

A simple query (“natural language generation”) in the VLO returns you interesting resources used for NLG:

Syntax Maker - The NLG tool for Finnish

(Part of B2SHARE: CLARIN)

⊞ Syntax maker is the **natural language generation** tool for generating syntactically correct sentences in Finnish automatically. The tool is especially useful in the case of Finnish which has such a high diversity in its morphosyntax. All you need to know are the lemmas and their parts-of-speech and syntax maker will take…

https://vlo.clarin.eu/

VLO and NLG resources

A simple query (“natural language generation”) in the VLO returns you interesting resources used for NLG:

Czech restaurant information dataset for NLG

(Part of LINDAT / CLARIAH-CZ Data & Tools)

⊞ This is a dataset for **natural language generation** (NLG) in task-oriented spoken dialogue systems with Czech as the target **language**. It originated as a translation of the English San Francisco Restaurants dataset by Wen et al. (2015). It includes input dialogue acts and the corresponding output **natural language** parap…

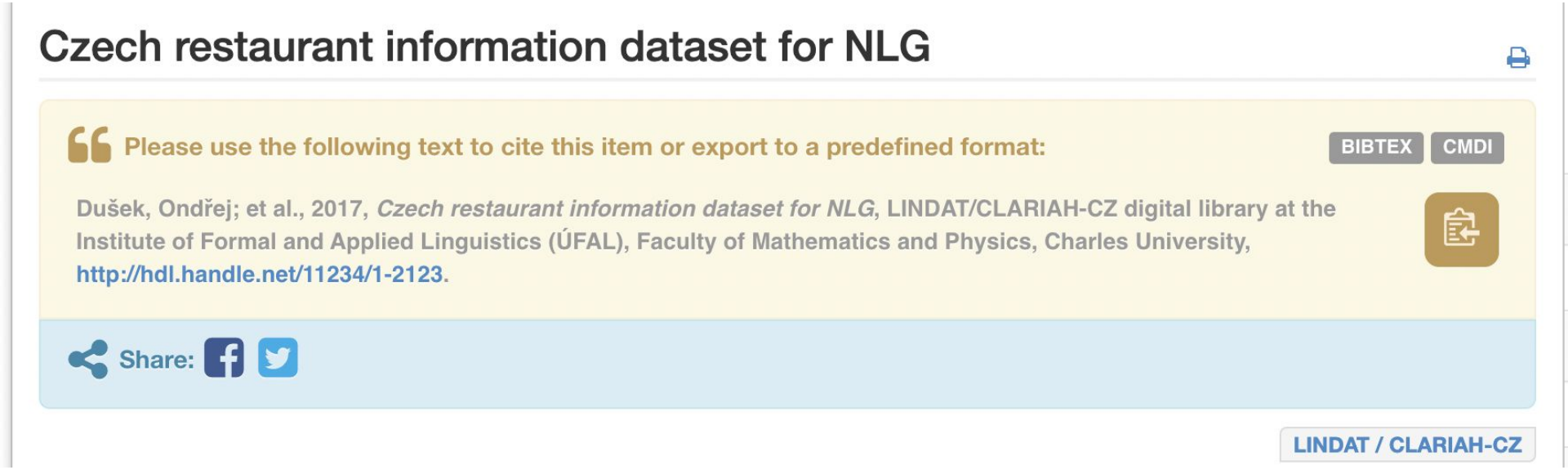
Czech

Landing page for this record

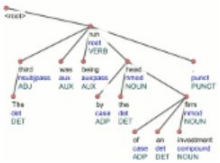
https://vlo.clarin.eu/

VLO and NLG resources

A click will take you to the **landing page**, that is to say the **actual repository that hosts the resource,** where you can access it.



UDPipe



Authors:
Milan Straka, Jana Straková

Description:

UDPipe is an trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given only annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary, as a library for C++, Python, Perl, Java, C#, and as a web service. UDPipe is a free software under Mozilla Public License 2.0 and the linguistic models are free for non-commercial use and distributed under CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions.

Project home Run

Language Resource Switchboard Upload Tool Inventory Help

Tool Inventory

Constituency Parsing

- WebLicht Const Parsing DE
- WebLicht Const Parsing EN

Coreference Resolution

- Concraft -> Bartek

Dependency Parsing

- Concraft -> DependencyParser
- MaltParser
- Spacy (hosted by D4Science) - DE
- Spacy (hosted by D4Science) - EN
- UDPipe
- WebLicht Dep Parsing DE
- WebLicht Dep Parsing EN

switchboard.clarin.eu

CLARIN Resource families

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

Spoken corpora in the CLARIN infrastructure

Corpora with transcriptions and audio recordings

Corpus	Language	Description	Availability
Arabic Speech Corpus **Licence:** CC BY 4.0	Arabic	The corpus is available for download from a dedicated webpage. For a relevant publication, see Halabi (2016).	Download
DIALEKT v1: dialectal corpus with multi-tier transcription **Size:** 100,000 words **Annotation:** orthographically and phonetically (dialect features) transcribed, MSD-tagged, lemmatised **Licence:** Academic Licence Agreement for Czech National Corpus Data	Czech	This corpus contains traditional dialectological material, mostly unprepared monologue-type speech. The corpus is available download (upon request) and through the concordancer KonText. For a related publication, see Komrsková et al. (2018).	Concordancer Download

https://www.clarin.eu/resource-families

CLARIN and Open Science

- Promoting the sharing and re-use of data through sustainable data registries
- All integrated datasets available in open access for research purposes
- Adherence to the FAIR data principles
  - Findable, Accessible, Interoperable, Re-usable
  - Interoperability through a common metadata framework
- Promotion of responsible data science
- Support for linguistic diversity
  - Data covering more than 1500 languages
  - Tools for many languages
  - Language resources in all modalities
- Strengthening the support for professional SSH researchers (> 500.000 in Europe)

CLARIN: Towards FAIR and Responsible Data Science Using Language Resources." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018, 3259-3264.

Exercise 1 - Finding and processing data

- Go to clarin.eu
- Find the VLO
- Search for texts by Robert Louis **Stevenson**
  - what can you find?
  - which format is the corpus in?
  - where are they hosted?
- On the Links tab, use the three dots **(...)** to activate the Switchboard.
  - Explore with **Voyant** or
  - Process with **UDpipe**

CLARIN for knowledge sharing

CLARIN Knowledge Infrastructure

Tour de CLARIN

Funding

Video Channels

Best-Practice Papers

https://www.clarin.eu/content/knowledge-infrastructure

Knowledge Centres

Knowledge Centres

CLARIN Knowledge Infrastructure

CLARIN Knowledge Centres (K-centres) are a cornerstone of the CLARIN Knowledge Infrastructure (KI), one of the main components ensuring a continuous transfer of knowledge between all players involved in the construction, operation and use of the infrastructure. The mission of the CLARIN KI is to ensure that the available knowledge and expertise does not exist as a fragmented collection of unconnected bits and pieces, but is made accessible in an organised way to both the CLARIN community and the social sciences and humanities research community more widely.

The Role of K-Centres

The focus of CLARIN is on language resources (in all modalities, from all regions and with any topical orientation) and K-centres serve researchers and educators from any discipline where language plays one of its many roles, ranging from object of study, a means of communication or expression, a means to store and extract information, object of learning or teaching activities, to training source for data-driven analytics, and many others. K-centres share their knowledge and expertise on one or more aspects of the domain covered by the CLARIN infrastructure and can be mostly found in CLARIN countries, but also exist elsewhere, and they all have a virtual presence.

Areas of K-Centre Expertise

K-centres all have their own specific areas of expertise, which can belong to many different categories, such as

https://www.clarin.eu/content/knowledge-centres

The Knowledge Centres of CLARIN can be contacted through their

Help Desks

Knowledge Centres

List of all 22 CLARIN K-centres with expertise in specific linguistic topis

Click on the full name of the K-centre to go to its landing page, and click on the acronym to see its full organisation details

ACE	**CLARIN Knowledge Centre for Atypical Communication Expertise**
Areas of competence	Atypical communication encompasses language and speech as encountered during (second) language acquisition and development, and in language disorders, but also more broadly in bilingual language development and in sign language. ACE is specialised in this type of research and concomitant infrastructural issues related to data acquisition, processing and sharing, which is typically highly characterised by sensitivity issues. For data storage and access the centre collaborates with MPI's TLA (The Language Archive) which is a CLARIN B Centre and also based in Nijmegen.
Audiences served	- linguists; - psychologists; - neuroscientists; - computer scientists; - speech and language therapists; - education specialists
Types of services	- how-to documents; - access to document templates; - Access to data; - Depositing; - FAQ; - Helpdesk; - Technical support
Is portal for language(s)	-
Other languages covered	-
Modalities covered	- Audio: speech; - Text; - Video: sign language
Linguistic topics	- Language acquisition (L1 and L2); - language disorders; - Language learning
Language processing	-
Data types	-
Resource families	- Spoken corpora; - Manually annotated corpora; - Multimodal corpora
Generic topics	- Critical Data Management; - Legal and ethical issues
Other keywords	- Language acquisition; - sign language; - language pathologies
Tour de CLARIN	Introduction Interview

https://www.clarin.eu/content/knowledge-centres

CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages (SAFMORIL)

SAFMORIL brings together researchers and developers in the area of computational morphology and its application during language processing. The focus of SAFMORIL is actual, working systems and frameworks based on linguistic principles and providing linguistically motivated analyses and/or generation on the basis of linguistic categories. Such systems are relevant in particular for languages with rich morphologies, e.g. Nordic and Baltic languages (such as Finnish, Swedish, Norwegian, Latvian, Lithuanian as well as the Sámi languages) and more generally Fenno-Ugric languages, Inuit languages, Canadian First Nation languages and Babylonian languages.

We offer online courses for developing and teaching morphologies, tokenizers and spell-checkers, a repository for storing morphologies, and an environment for creating tokenizers and spell-checkers. SAFMORIL serves linguists and computational linguists developing and adapting morphologies as well as digital humanities scholars and computer scientists processing language data.

If you have any questions, contact us via the SAFMORIL **Helpdesk (safmoril@kielipankki.fi)**.

CLARIN K CENTRE

CLARIN

FIN-CLARIN

https://www.kielipankki.fi/safmoril/

Impact Stories:
Showcases high-quality and innovative research that uses CLARIN tools and resources.

IceTaboo: Offensive Word Database with Commercial Application

The IceTaboo database can be used to flag contextually inappropriate words in texts, and is already being used as part of an automatic proofreading software by an Icelandic online news website.

Read more

https://www.clarin.eu/content/clarin-impact-stories

Video Lectures, training materials

Training Materials

Applied Language Technology
Author: Tuomo Hiippala
Faculty of Arts, University of Helsinki, Finland

Archilochus of Paros: Elegiac Fragments - XML Archive
Author: Anika Nicolosi and Beatrice Nava
University of Parma, Italy

Computational Morphology with HFST
Author: Erik Axelson
Faculty of Arts, University of Helsinki, Finland

GATE, an Open-Source Toolkit for Natural Language Processing
Author: Diana Maynard
Faculty of Engineering, University of Sheffield

Introduction to Digital Humanities
Author: Zuzana Neverilova
Faculty of Arts, Masaryk University, Czech Republic

Introduction to Speech Analysis
Author: Mietta Lennes
Faculty of Humanities, University of Helsinki, Finland

https://www.clarin.eu/content/teaching-clarin

Past CLARIN Cafés

TOPIC

CLARIN Café on Text+ : A New Research Data Initiative in Germany
Organised by Erhard Hinrichs (University of Tübingen), Thorsten Trippel (IDS Mannheim), and Andreas Witt (IDS Mannheim)

CLARIN Café: Towards Guidelines for Integrating CLARIN into Teaching
Organised by **Darja Fišer** (University of Ljubljana), Iulianna van der Lek CLARIN *ERIC*

CLARIN Café on Text and Data Mining Exceptions in the Directive on Copyright in the Digital Single Market

CLARIN Café: Interactive Q&A Session for newcomers in CLARIN from the SSH domain

https://www.clarin.eu/content/clarin-cafe

CLARIN Funding Hub

On this page you can find all CLARIN funding opportunities, as well as information about additional support that CLARIN can offer researchers preparing project proposals for EU-funded projects or organising virtual events. Below, you can see some recently approved, funded projects, along with a comprehensive list of all current and past projects funded by CLARIN.

CLARIN Calls

CLARIN offers funding to address strategic priorities that require international collaboration, exchange of expertise, training or mobility. If you would like to suggest a new topic for which funding may be beneficial, please get in touch.

https://www.clarin.eu/content/funding-opportunities

FAIR by design

Thanks to its network of **K-centres** and its **support instruments**, CLARIN not only allows researchers to share and find LRs, but also to create LRs that are

- FAIR by design
- interoperable
- comparable

CLARIN ParlaMint Project

Towards Comparable Parliamentary Corpora

17 european parliaments
including Senato della Repubblica italiana

Word list

Corpus: ParlaMint-IT 2.0 (Italian parli
Subcorpus: COVID

Reference corpus: ParlaMint-IT 2.0 (It
Reference subcorpus: Reference
Switch focus and reference (sub)corpu

Page 1 Go Next >

word	frequency	ParlaMint-IT 2
pandemia	9	
Covid-19	6	
Covid	6	
coronavirus	5	
lockdown	3	
mascherine	3	
recovery	3	
MES	6	
distanziamento	2	
tamponi	2	
fund	3	
contagi	2	
generation	1	
mascherina	1	
DPCM	1	
FIBP	1	
Azzolina	107	31.0
virus	910	263.9
sierologici	100	29.0

Word list

Corpus: ParlaMint-GB 2.0 (British parliament)
Subcorpus: COVID_GB

Reference corpus: ParlaMint-GB 2.0 (British parliament)
Reference subcorpus: MPregReference
Switch focus and reference (sub)corpus

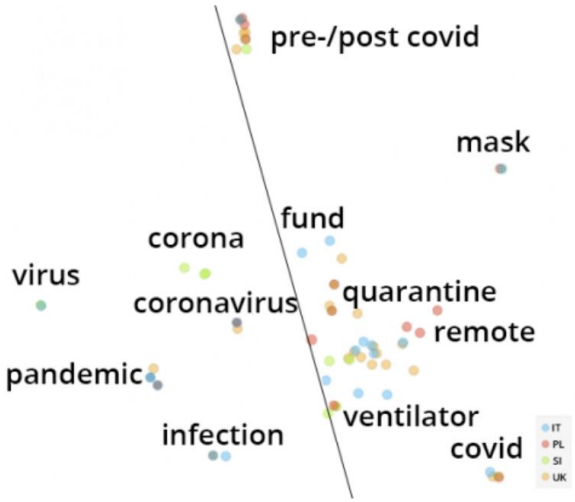
Page 1 Go Next >

word	frequency	frequency/mill	ParlaMint-GB 2.0 (British parliament) : COVID_GB	ParlaMint-GB 2.0 (British parliament) : MPregReference		Score
			frequency	frequency/mill		
covid	9,667	412.9	0	0.0		413.9
pandemic	10,341	441.7	1	0.1		402.4
coronavirus	5,698	243.4	0	0.0		244.4
Covid	5,676	242.4	0	0.0		243.4
lockdown	5,571	237.9	2	0.2		199.1
furlough	2,129	90.9	0	0.0		91.9
PPE	1,564	66.8	0	0.0		67.8
virus	6,009	256.7	40	4.0		51.5
distancing	1,853	79.1	6	0.6		50.1
CHIS	860	36.7	0	0.0		37.7
lockdowns	666	28.4	0	0.0		29.4
Coronavirus	652	27.8	0	0.0		28.8
SAGE	572	24.4	0	0.0		25.4
tracing	1,048	44.8	8	0.8		25.4
isolate	1,579	67.4	20	2.0		22.8
quarantine	834	35.6	7	0.7		21.5
masks	908	38.8	9	0.9		20.9
vaccine	4,856	207.4	91	9.1		20.6
trace	1,748	74.7	27	2.7		20.5
shielding	492	21.0	1	0.1		20.0
Virtual	397	17.0	0	0.0		18.0
furloughed	396	16.9	0	0.0		17.9
Trace	389	16.6	0	0.0		17.6
inaudible	370	15.8	0	0.0		16.8
COVID	363	15.5	0	0.0		16.5
infection	1,697	72.5	36	3.6		16.0
			1	0.0	30.9	
			208	7.6	30.9	
			1	0.0	28.9	

https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

Helsinki Digital Humanities Hackathon 2021: ‘Parliamentary Debates in COVID Times’

Contributors: Contributors: Isabella Calabretta, Courtney Dalton, Richard Griscom, Marta Kołczyńska, Matej Klemen, Kristina Pahor de Maiti, Ajda Pretnar Žagar, Ruben Ros



‘Compiling a corpus is already a big project, so being able to skip this step was a huge privilege. Also, knowing that the corpus was granted permission to be included in the CLARIN repository already gives you some idea of its quality.’
Kristina Pahor de Maiti

https://www.clarin.eu/impact-stories/helsinki-digital-humanities-hackathon-2021-parliamentary-debates-covid-times

Exercise 2 - ParlaMint

- Explore the ParlaMint corpora using NoSketch Engine (public): clarin.si/noske/index-en.html.
- Find the Italian corpus
  - Explore the **Corpus info -** how many subcorpora are there?
  - Create a wordlist for the COVID subcorpus
- Keyword extraction
  - extract **keywords** for the "COVID" subcorpus using the "REFERENCE" as **Reference (sub)corpus**

The CLARIN-IT national consortium

CLARIN-IT (2015-onwards)

CLARIN-IT

About | Governance | Consortium | Centres | Join | Access | Events | Initiatives | News | Home

the Italian Common Language Resources and Technology Infrastructure

English

ONLINE EVENTS

EUPORIA 2021 Webinar - Encoding a Critical Apparatus

07/12/2020 - Registration to follow the Webinar via Zoom

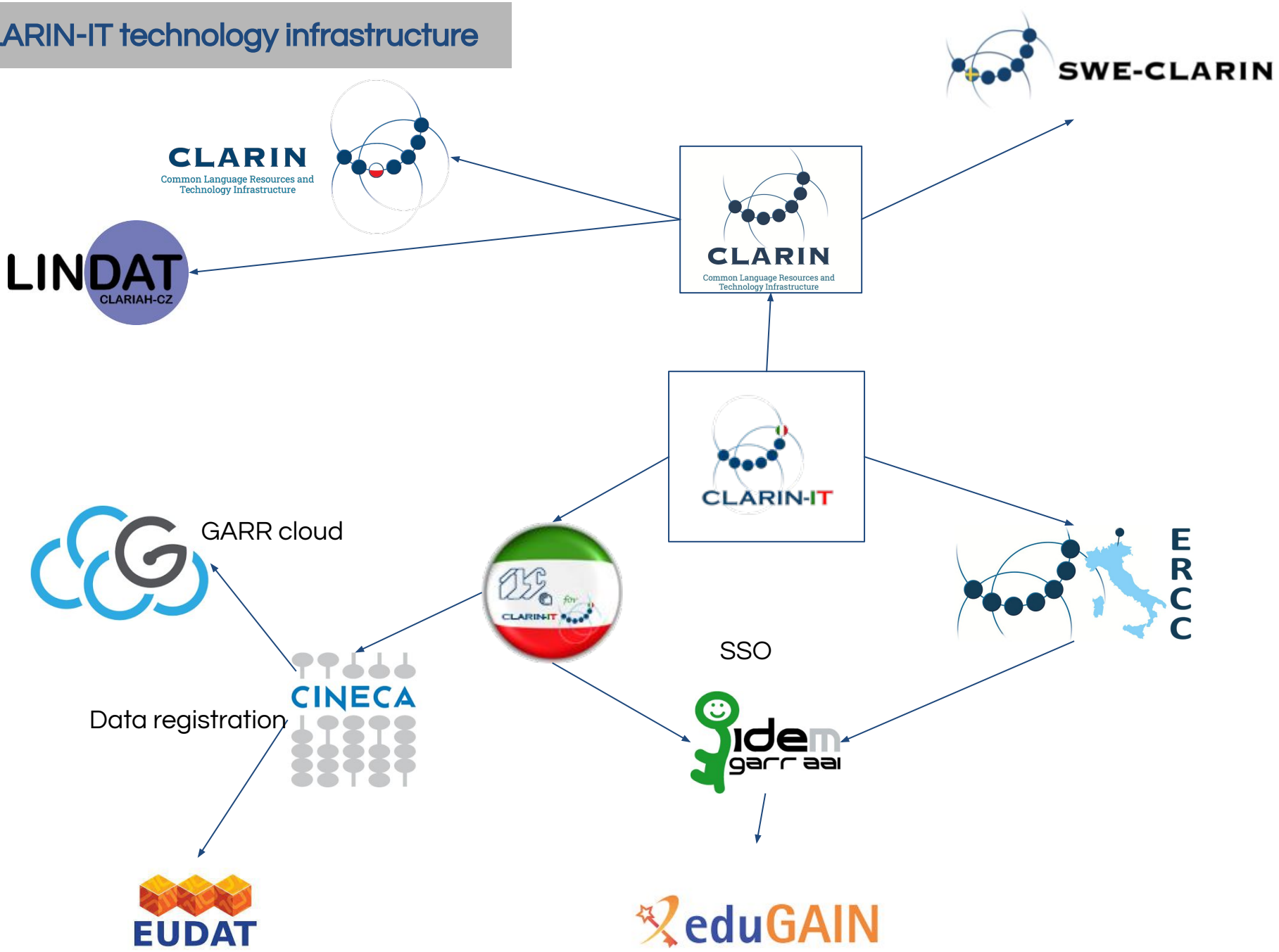
CLARIN-IT Consortium

- The Department of Education, Human Sciences and Intercultural Communication of the University of Siena
- The Department of Philology and Literary Criticism of the University of Siena
- The Eurac Research Association (Bolzano)
- The Bruno Kessler Foundation (Trento)
- The Archival and Bibliographical Superintendence of Tuscany (Firenze)
- The Department of Electrical Engineering and Information Technology and the Interdepartmental Research Center "URBAN/ECO" of the University of Naples Federico II
- The Catholic University of the Sacred Heart (Milano)
-
- The University of Parma (Parma) has started the membership procedure.

CLARIN-IT Research Topics

- **Resources and Tools for the Italian Language**
  - create new resources by enriching existing corpora
  - lexical datasets with Linked Open Data
  - specialized corpora for computer-mediated communication.
  - natural language processing and analysis tools, offered as web services and integrated into Weblicht.
- **Resources for Regional Languages and Multilingual corpora**
  - learner corpus for German, Italian and Czech,
- **Speech Archives**
  - Grafo, Caterina Bueno Archive
- **Digital Classics**
  - resources for Ancient Greek and Latin (LOD version of the TEI-dict Perseus Liddell-Scott Jones dictionary)
  - Italian Latinity of the Middle Ages
  - digital editions of ancient fragmentary texts

CLARIN-IT technology infrastructure



ILC4CLARIN


CLARIN
B CENTRE
CLARIN-IT
CLARIN
Chi siamo
Organizzazione
Repository
Servizi
Eventi CLARIN
ILC4CLARIN
REPOSITORY
Easy to be found | Easy to be cited

ILC4CLARIN: data types

CLARIN-IT: i servizi

Editing lessicale

LexO

Editor Web collaborativo utilizzato per la creazione e la gestione di risorse lessicali e terminologiche (multilingue) come risorse di dati collegati.

Approfondisci

Annotazione/analisi del testo

Freeling IT

Eseguibile di freeling solo per la lingua italiana, up to tagged.

Prova questo strumento

DeSR

Parser shift/reduce di dipendenze per la lingua italiana.

Prova questo strumento

LTFW (Linguistic Tools For Weblicht)

Porting Java del tokenizzatore basato su perl sviluppato all'interno del progetto OpeNER.

Prova questo strumento

Merging

LMF Merger

Merger LMF di due lessici italiani.

Approfondisci

LMF ML Merger

Merger LMF multilivello.

Approfondisci

Multiword Extractor: Average Frequency Filter

Estrattore di parole multiple che applica un filtro di frequenza media su MWE.

Prova questo strumento

Multiword Extractor: Merge Overlapping Strings

Estrattore di parole multiple che applica un filtro che unisce stringhe sovrapposte.

Prova questo strumento

Multiword Extractor: Remove Substrings

Estrattore di parole multiple che applica un filtro che rimuove le sottostringhe.

Prova questo strumento

Multiword Extractor: Reorder

Estrattore di parole multiple che applica un filtro che riordina i dati ExtractorMW per coppia_frequenza.

Prova questo strumento

Lexical services

ItalWordNet

Parola: Invia Mostra tutte le relazioni

---

casa, Nome

- [1] - edificio o parte di esso in cui si abita;
  (*abitazione* [1], *casa* [1], *dimora* [2], *magione* [1], *ostello* [2], *tetto* [4])

- [2] - edificio con particolari funzioni.
  (*casa* [2])

- [3] - dinastia, casa regnante; l'insieme dei sovrani, appartenenti a una stessa famiglia, che si succedono sul trono
  (*casa* [3], *casa regnante* [1], *dinastia* [1])
  "la casa di Savoia non ha regnato a lungo sull'Italia"

- [4] - la famiglia alla quale si appartiene
  (*casa* [4])
  "sta pensando di mettere su casa"

- [5] - ditta, compagnia, in particolare nel settore dell'editoria e della moda
  (*casa* [5])
  "casa editrice"

- [6] - casella nel gioco degli scacchi
  (*casa* [6])

NLP services

URL del servizio: freeling_it (WSDL)

Come usare il servizio

Prova questo servizio attraverso il form sottostante.

Run service

Inputs **Report**

input as URL / direct data or local file

Choose file No file chosen

language it.cfg

mandatory
optional

multiword yes no

ner --Use Default--

output_format --Use Default--

Reset fields

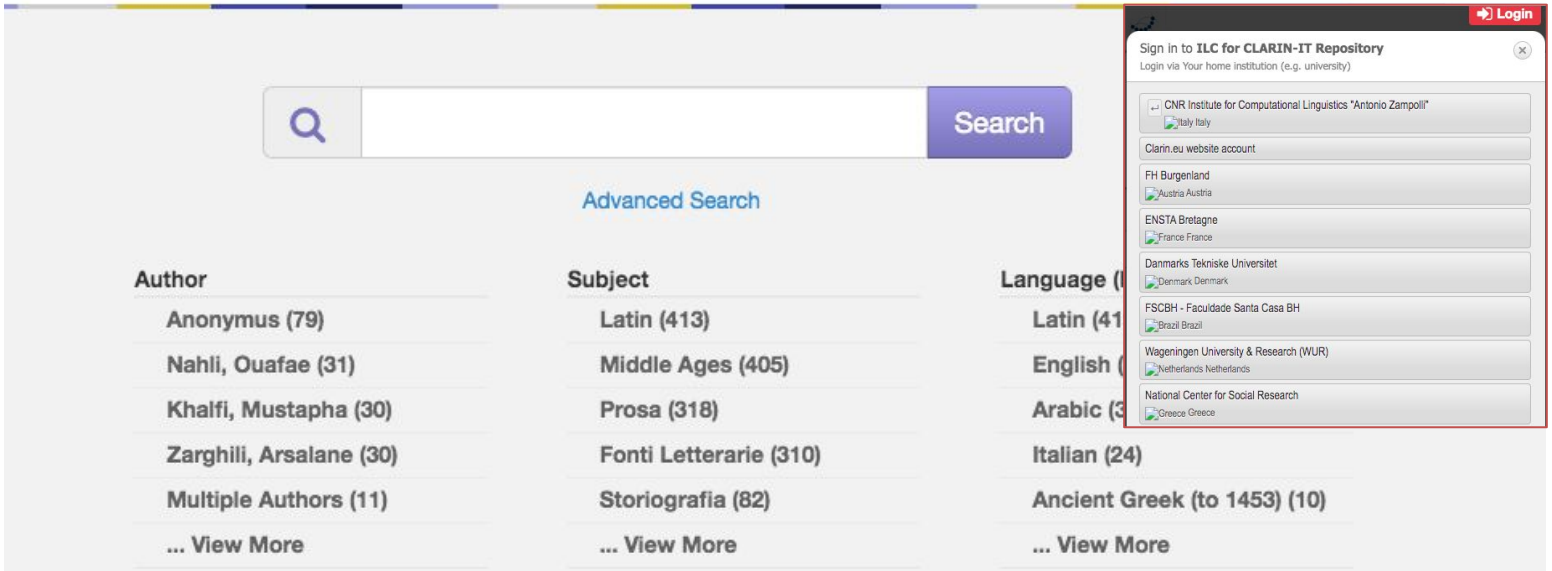
ILC4CLARIN: deposit

Deposit Free and Safe

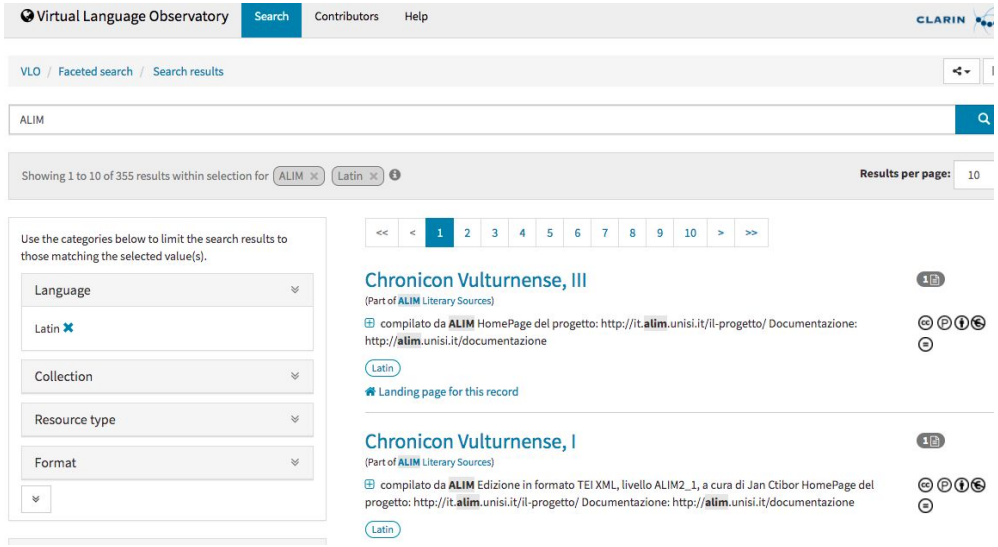
License of your Choice (Open licenses encouraged)

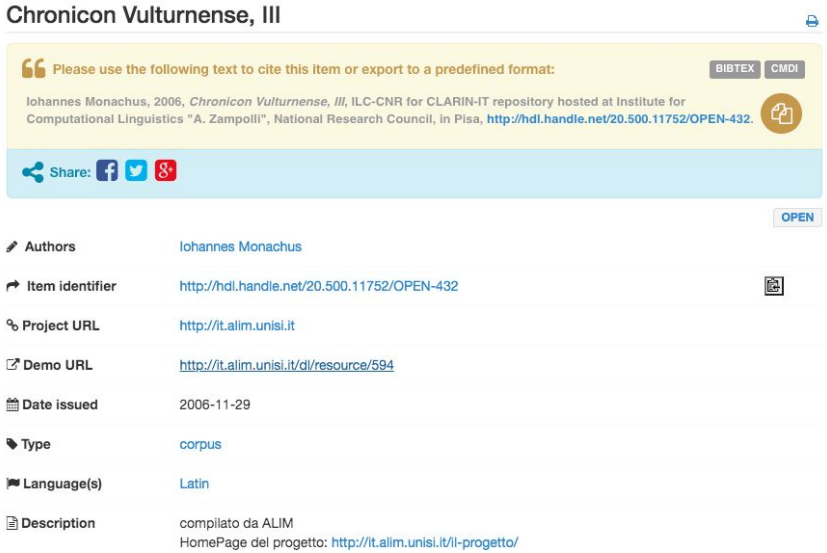
Easy to Find

Easy to Cite



ILC4CLARIN: findability, access





ILC4CLARIN: interoperability,reuse

Resources

nn_nrhz120_1848.TEI-P5.xml 106.14 KiB

Mediatype: text/plain

Language: German

Matching Tools

Group by task ☑ Search for tool

- Constituency Parsing
  - WebLicht > Open — WebLicht Const Parsing DE
- Dependency Parsing
  - D4Science Infrastructure > Open — Spacy (hosted by D4Science) - DE
  - LINDAT > Open — UDPipe
  - WebLicht > Open — WebLicht Dep Parsing DE
- Distant Reading
  - > Open — Voyant Tools

K-Centres & DH: DiPText-KC

CLARIN K CENTRE DiPText-KC ABOUT PARTNERS PEOPLE KNOWLEDGE HELPDESK EVENTS NEWS CONTACT

DiPText-KC

CLARIN Knowledge Centre for Digital and Public Textual Scholarship

DiPText-KC offers expertise on methods, data, instruments and technologies relevant in the field of Philological and Literary Studies, History, Art History and Cultural Heritage.

Its actions aim at:

- sharing information with scholars and students about the state of the art in digital scholarly editing and text annotation through domain-specific languages;
- supporting scholars and students in the creation and publication of digital scholarly editions and resources;
- organizing training activities (for instance webinars, workshops and summer schools).

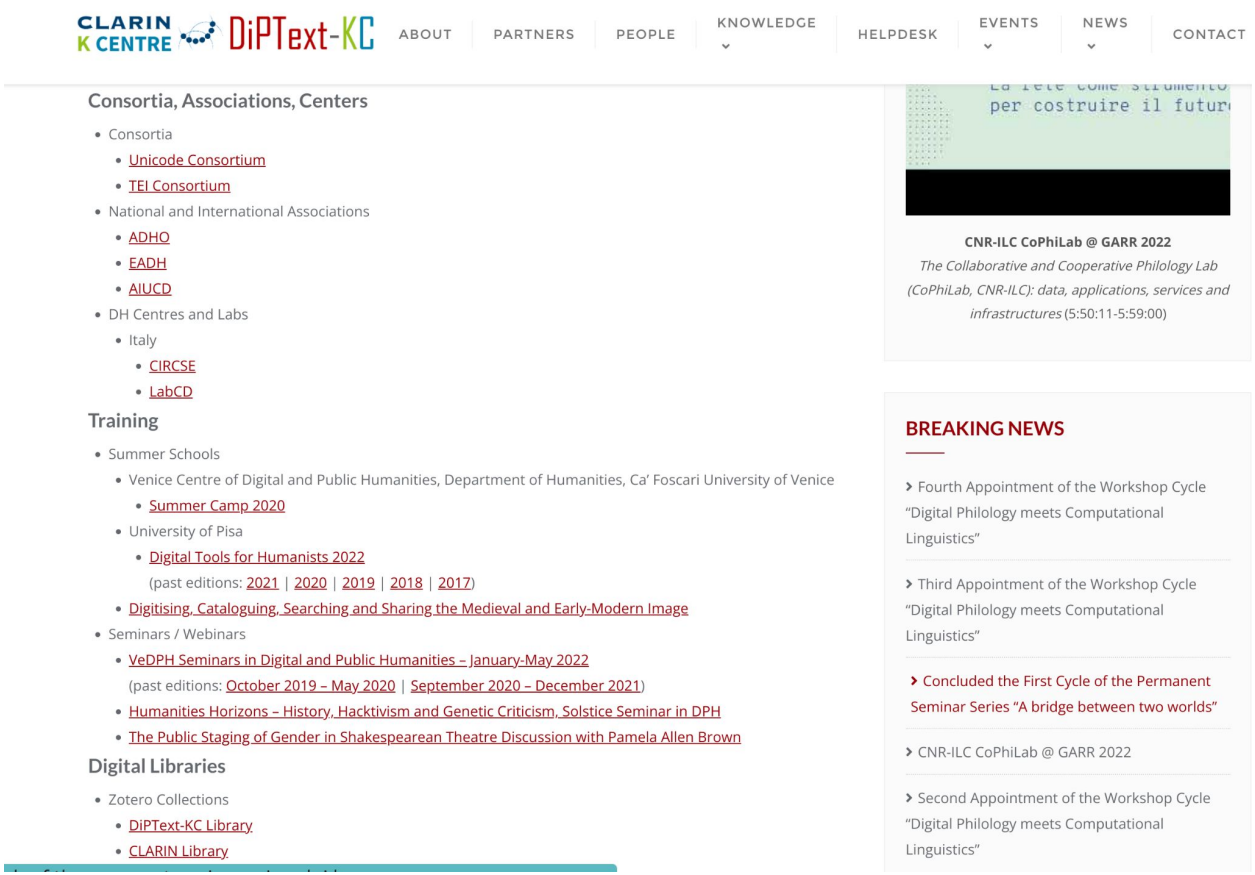
DiPText-KC is one of the Centres of CLARIN-IT, the Italian node of CLARIN (Common Language Resources and Technology Infrastructure), a digital infrastructure of pan-European interest identified by ESFRI (European Strategy Forum on Research Infrastructures) and classified as a Landmark Research Infrastucture for the Social Sciences and Humanities (ESFRI Landmarks SSH RI).

HIGHLIGHTS

The Digital and Public Textual Scholarship Knowledge Centre is focused on **digital philology**

https://diptext-kc.clarin-it.it

Activities of the DiPText-KC



The Digital and Public Textual Scholarship Knowledge Centre keeps you informed on Consortia, Associations, Centres, Training Schools, and Digital Libraries relevant for digital philologists

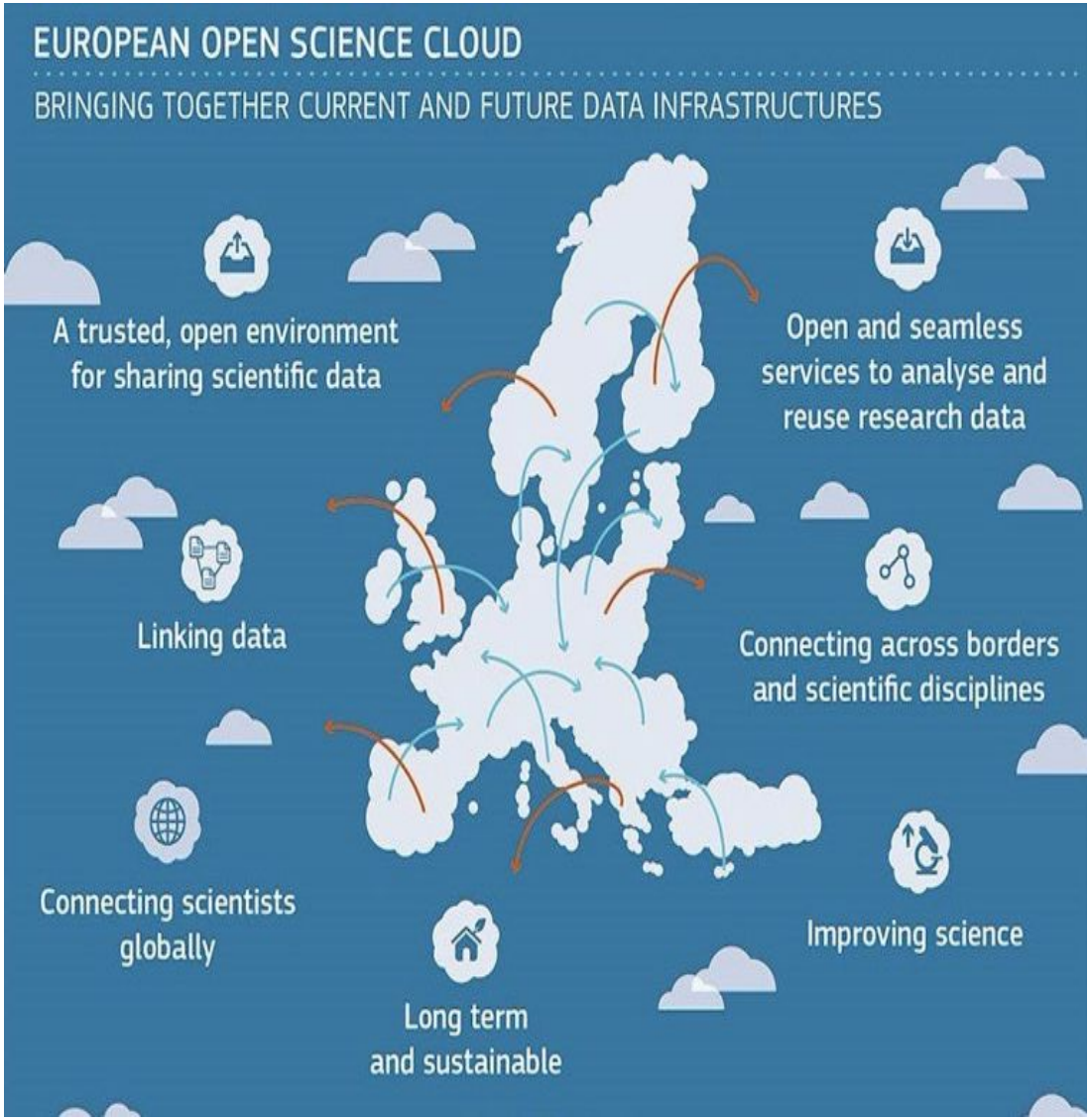
https://diptext-kc.clarin-it.it

Exercise 3 - Explore ILC4CLARIN

- Find the ILC4CLARIN repository
  - Which is the most represented language in terms of records?
  - What kind of data can you find in Arabic?
  - How do you cite CophiWordNet?
- Try to log in with your institutional identifier, using the Login function (top right)

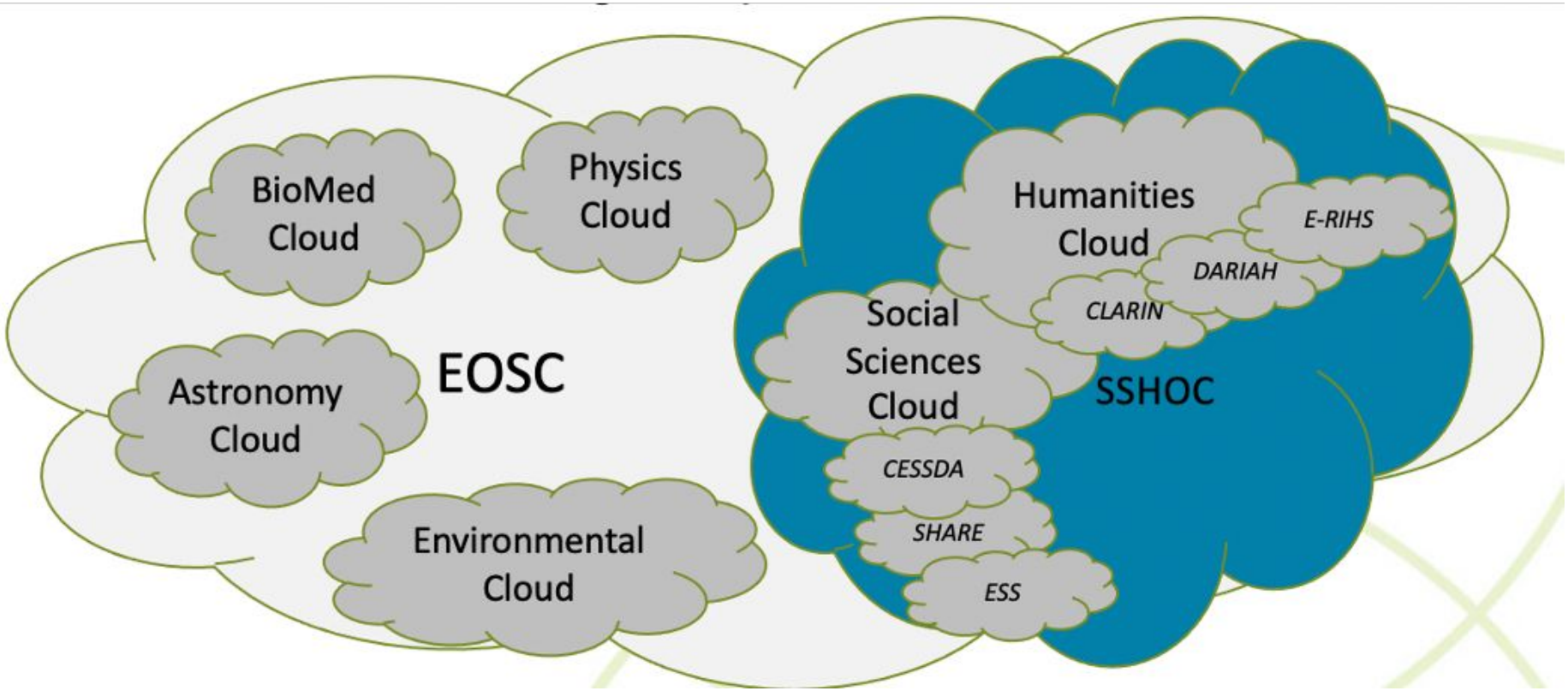
A Cluster of SSH Research Infrastructures

EOSC: a challenge



- The European Union has launched the European Open Science Cloud as a supporting landscape to foster open science and open innovation
- The EOSC is aimed at removing technical, policy and human barriers, leading to knowledge creation and economic prosperity in Europe.

The SSH Open Cluster



Social Sciences & Humanities Open Cloud

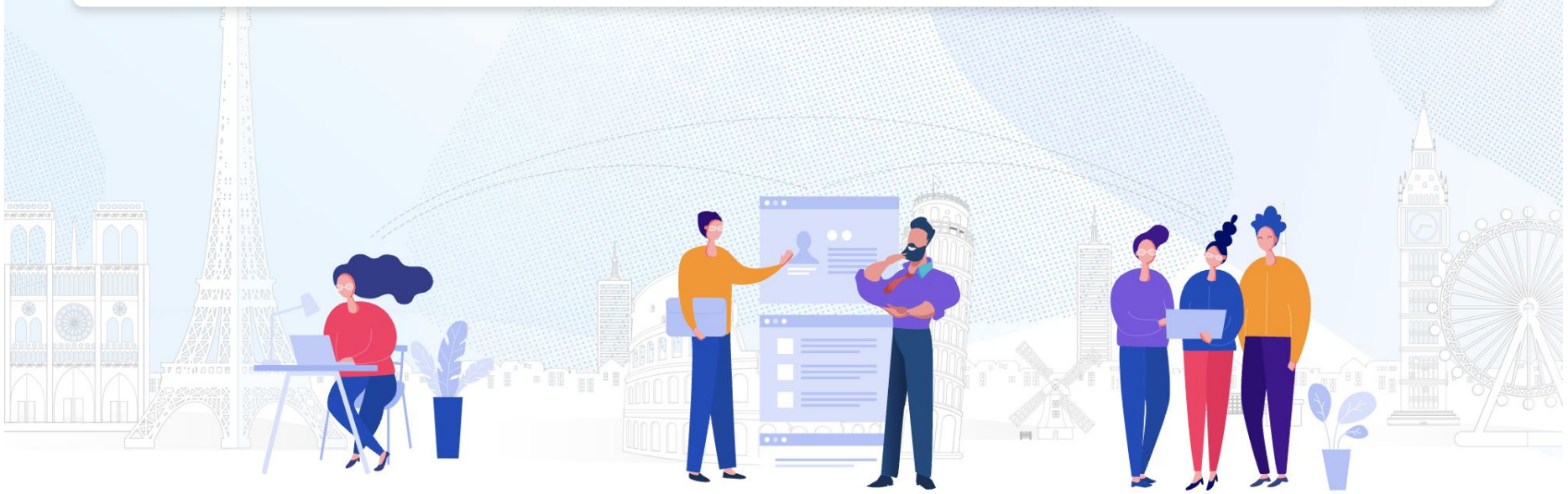
https://www.sshopencloud.eu/

Social Sciences & Humanities Open Marketplace

Discover new and contextualised resources for your research in Social Sciences and Humanities: tools, services, training materials, workflows and datasets. Read more…

All categories

Search



Search results (5361)

Refine your search

Clear filters

Sort by Relevance

Previous 1 of 269 Next

CATEGORIES

- Tools & services 1731
- Training materials 340
- Publications 2953
- Datasets 310
- Workflows 27

ACTIVITIES

- Analyzing 641
- Visual Analysis 314

COLLADA Schema Version 1.5.0

No description provided.

Read more

COLLADA -- Digital Asset Schema Release 1.5.0 - Specification

This document describes the COLLADA schema. COLLADA is a COLLAborative Design Activity that defines an XML-based schema to enable 3D authoring applications to freely exchange digital assets without loss of information, enabling multiple software packages to be combined into extremely…

Read more

Exercise 4 - The SSHOC Marketplace

- Explore the training materials on the SSHOC Marketplance
- Can you find anything that can be useful for your research?
- If so, can you tell us why?

Facilitating Data Reuse

- Make your data **FAIR**
- Depositing in trusted **repositories**
- Acknowledge the effort made by others by properly **citing datasets** in your publications
- Work towards building **interoperable services**

CLARIN and its national nodes can help you!

CLARIN for researchers

- Contact CLARIN for help with your research
- Visit this page
  - https://www.clarin.eu/content/clarin-researchers
- Participate in CLARIN (virtual) events
  - https://www.clarin.eu/events
- Tour de CLARIN
  - https://www.clarin.eu/Tour-de-CLARIN
- Use CLARIN Training and videolectures

Contacts

- CLARIN ERIC
  - Newsletter https://www.clarin.eu/news
  - @CLARINERIC
- CLARIN IT
  - www.clarin-it.it
  - @CLARIN_IT
  - coordination@clarin-it.it