Linguistic Linked Open Data for Humanists

Anas Fahad Khan, Giulia Pedonese, Michele Mallia

Lisbon Summer School in Linguistics, July 1-5, 2024

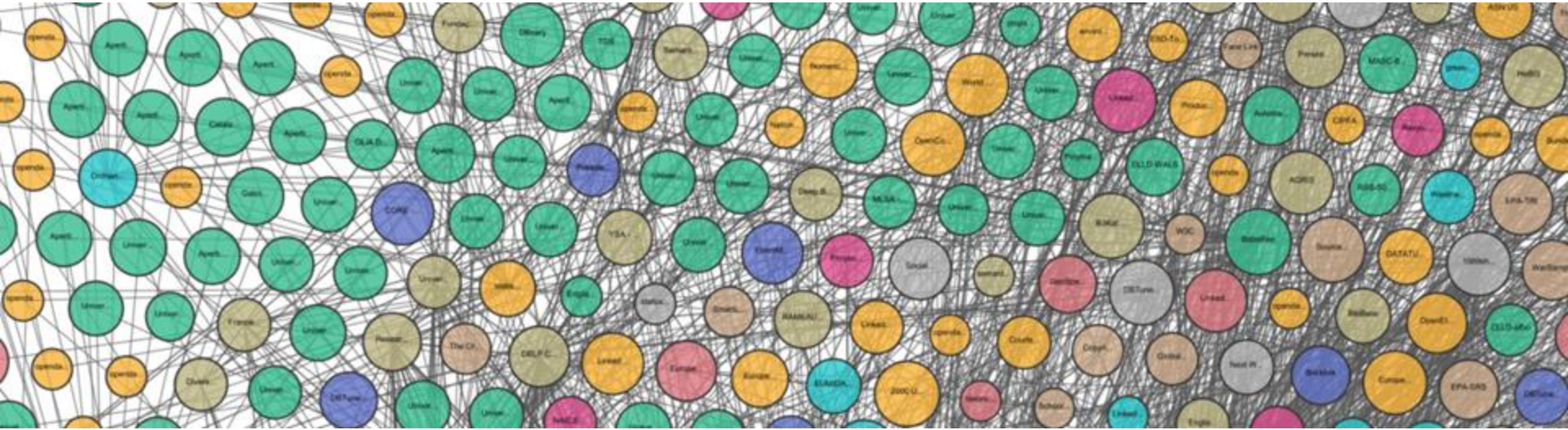
Funded by the European Union NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Consiglio Nazionale delle Ricerche

Language resources, FAIR principles and Linked Open Data



The Linked Open Data Cloud from lod-cloud.net

Anas Fahad Khan
Giulia Pedonese
Michele Mallia

Sources

- Steven Krauwer, Bente Maegaard. "CLARIN – How It Started". CLARIN: The Infrastructure for Language Resources, edited by Darja Fišer and Andreas Witt, Berlin, Boston: De Gruyter, 2022, pp. 1-30. https://doi.org/10.1515/9783110767377-001
- "FAIR Guiding Principles for scientific data management and stewardship", 2016 https://www.go-fair.org/fair-principles/
- Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. "CLARIN's Key Resource Families". In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).
- Iulianna van der Lek, Darja Fišer. (2023). "Introduction to Language Data: Standards and Repositories". In UPSKILLS Learning Content. https://upskillsproject.eu/project/standards_repositories/. CC BY 4.0.

This presentation uses pictures from https://www.flaticon.com/

Index

- What are Language Resource and what they are useful for
- Discovering Language Resources
- The CLARIN infrastructure
- Examples of Language Resources in CLARIN
- Try it yourself: the Virtual Language Observatory
- How is it possible? The FAIR principles
- Metadata and Licenses
- FAIR principles and the LOD paradigm: differences and common goals

What are Language Resources?

A new concept defined as follows: «The term Language Resource (LR) refers to **a set of speech or language data and descriptions in machine-readable form**.

They are used for **building, improving or evaluating natural language** (human language) and **speech algorithms** or systems, and increasingly, for **machine learning**. They are also used as core resources for the software localisation and language services industries, for language and translation studies, electronic publishing, international transactions, subject-area specialists and end-users.»

Source: ELRA/ELDA

What are Language Resources?

- The terms **linguistic resources** and **language resources** are often used interchangeably
- The terms were **first used by Antonio Zampolli** in his paper *Towards reusable linguistic resources* at the EACL conference in 1991
- In 1992, the European Commission published **Danzin's report Towards a European Language Infrastructure** in which LRs were for the first time politically acknowledged as playing an important role for research, the language industry and Europe in general

References:
A. Zampolli. 1991. Towards Reusable Linguistic Resources. In Fifth Conference of the European Chapter of the Association for Computational Linguistics, Berlin, Germany. Association for Computational Linguistics.

Danzin, A (1992) Towards a European language infrastructure. Report by A. Danzin and the Strategic Planning Study Group for the Commission of the European Communities (DG XIII). [EU Commission - Working Document]

Typology of Language Resources

Language resources are typically divided into categories depending on the kind of content they include:

- **Textual resources** > written and spoken corpora
- **Lexical resources** > lexica, dictionaries and terminological databases
- **NLP Tools** > lemmatisers, PoS taggers, parsers ecc.

Language resources do not only include data, but also **data description, or metadata**, enriching data with additional information, such as structural division into books or linguistic traits like PoS tags and syntactic functions

Why metadata are as important as data

Metadata are data that provide information about other data. They summarise basic information about data, making them findable and reusable.
Some of the most used categories of metadata are:

Descriptive > author, content, name of the dataset
Structural > how data are classified, their format, ecc.
Administrative > licensing and management information
Relationship > explaining how the dataset relate to other information

https://www.ontotext.com/knowledgehub/fundamentals/metadata-fundamental/

SSHOC Multilingual Data Stewardship Terminology

ILC

Authors	Frontini, Francesca ; Gamba, Federica ; Monachini, Monica and Broeder, Daan
Item identifier	http://hdl.handle.net/20.500.11752/ILC-567
Project URL	https://www.sshopencloud.eu/
Demo URL	https://vocabs.sshopencloud.eu/vocabularies/sshocterm/
Date issued	2021-12-31
Type	lexicalConceptualResource, text
Size	210 concepts
Language(s)	Dutch , English , French , German , Italian , Modern Greek (1453-) , Slovenian
Description	The SSHOC Multilingual Data Stewardship Terminology is a multilingual terminology that collects terms specific to the domain of Data Stewardship, as well as their definitions. A list of domain-specific terms was automatically extracted from a corpus pertaining to the domain of Data Stewardship and Curation, validated by domain experts, assigned a definition, and linked to other existing terminologies (Loterre Open Science Thesaurus, terms4FAIRskills, Linked Open Vocabularies, ISO terms and definitions). Each term-definition pair was then automatically translated into multiple languages (Dutch, French, German, Greek, Italian, Slovenian) by employing Deep-L. The Multilingual Data Stewardship Terminology thus consists of 210 concepts available in Dutch, French, German, Greek, Italian, Slovenian. This resource was created within the frame of the SSHOC (Social Sciences and Humanities Open Cloud) project (H2020-INFRAEOSC-2018-2-823782). It is the result of the work of Task 3.1.2 "extraction of terminology from technical documentation about standards and interoperability", as described in D3.9, carried out jointly by ILC-CNR and CLARIN ERIC.

What are Language Resources useful for?

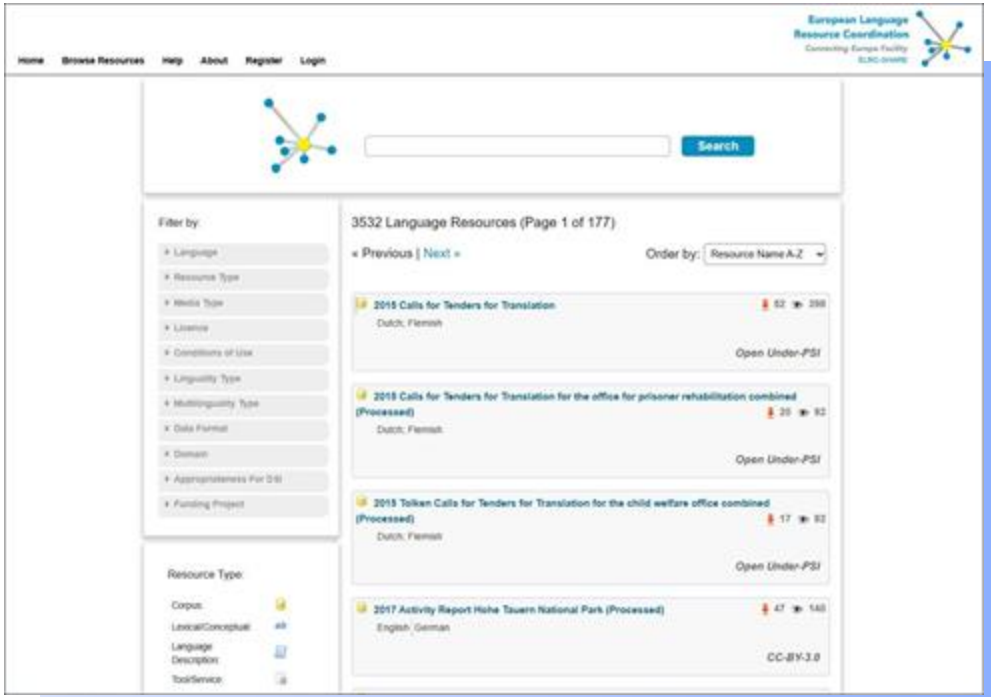
«**Language Resources** (LRs) are **raw material for Language Technologies** (LTs) **development and upgrading**, the medium for conveying information and knowledge (if possible in the most efficient and effective way), the content for developing culture and civilizing societies. LRs in combination with LTs **have tremendously changed the user experience and interactive possibilities of apps, tools and systems**, and public media over the years. Thus, LRs – in combination with LTs – have a huge economic and societal impact» Source: LT Innovate

NB: **Language Technologies** are computational methods, computer programs and electronic devices that are specialised for analysing, producing or modifying text and speech, e.g. spell checkers, machine translation, speech synthesis, question and answering ecc.

Language Resources and Repositories

To view different types of language resources and understand the way they are described (the metadata model), take a look at the [ELRC-SHARE repository]

“We define a **repository** as a service operated by research organizations, where research materials are stored, managed and made accessible.” Source: [DataCite]



Practice

Exercise 1

- How does ELRC-SHARE classify language resources?
- What resource types did you find?
- Take a closer look at each type of resource category and see how they are further categorised.
- How many media types are available?

Exercise icons created by Freepik - Flaticon

Examples of Language Resources

While there are many language resources available on the Web, we will show some examples from **CLARIN Resource Families** because these resources are curated and contain rich metadata and descriptions, such as size, text sources, time periods, annotations and licences.

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

The overviews have been compiled by Darja Fišer and Jakob Lenardič.

Reference: Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Corpora

- Computer-Mediated Communication Corpora
- Corpora of Academic Texts
- Historical Corpora
- L2 Learner Corpora
- Legal Corpora
- Literary Corpora
- Manually Annotated Corpora
- Multimodal Corpora
- Newspaper Corpora
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources

- Language Models
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

Parallel Corpora

Parallel corpora (also called translation corpora) contain **source texts in a given language which have been aligned with translations in another language.** They are often used not only for research in translation studies and **contrastive linguistics**, but also in **translator training**, foreign language **teaching**, bilingual **lexicography**, **terminology extraction**, **computer-aided translation**, training of **machine translation** systems and **cross-lingual information retrieval**.

You can find parallel corpora in the CLARIN Resource Families and OPUS.

Tools to query/analyse bilingual (and multilingual) parallel corpora

- Sketch Engine (commercial tool but offers flexible licences for academia)
- NoSketch (open-source, limited version of the commercial variant)
- AntConc (a free toolkit that can be downloaded on your local device)
- ParaConc (a commercial bilingual/multilingual concordancer that includes semi-automatic alignment, parallel searches, collocate extraction)

COMPARA

This page in English

Resultados da pesquisa

Voltar

Imprimir

Procura: **cat** Pedido de: **concordância em contexto**. Direcção da pesquisa: **De inglês para português**. Resultados: **62** ocorrências . Expressão de pesquisa: **"cat"**

Descrição do corpus usado nesta procura: **1435926** palavras portuguesas, **1542762** palavras inglesas, **97723** unidades de alinhamento.

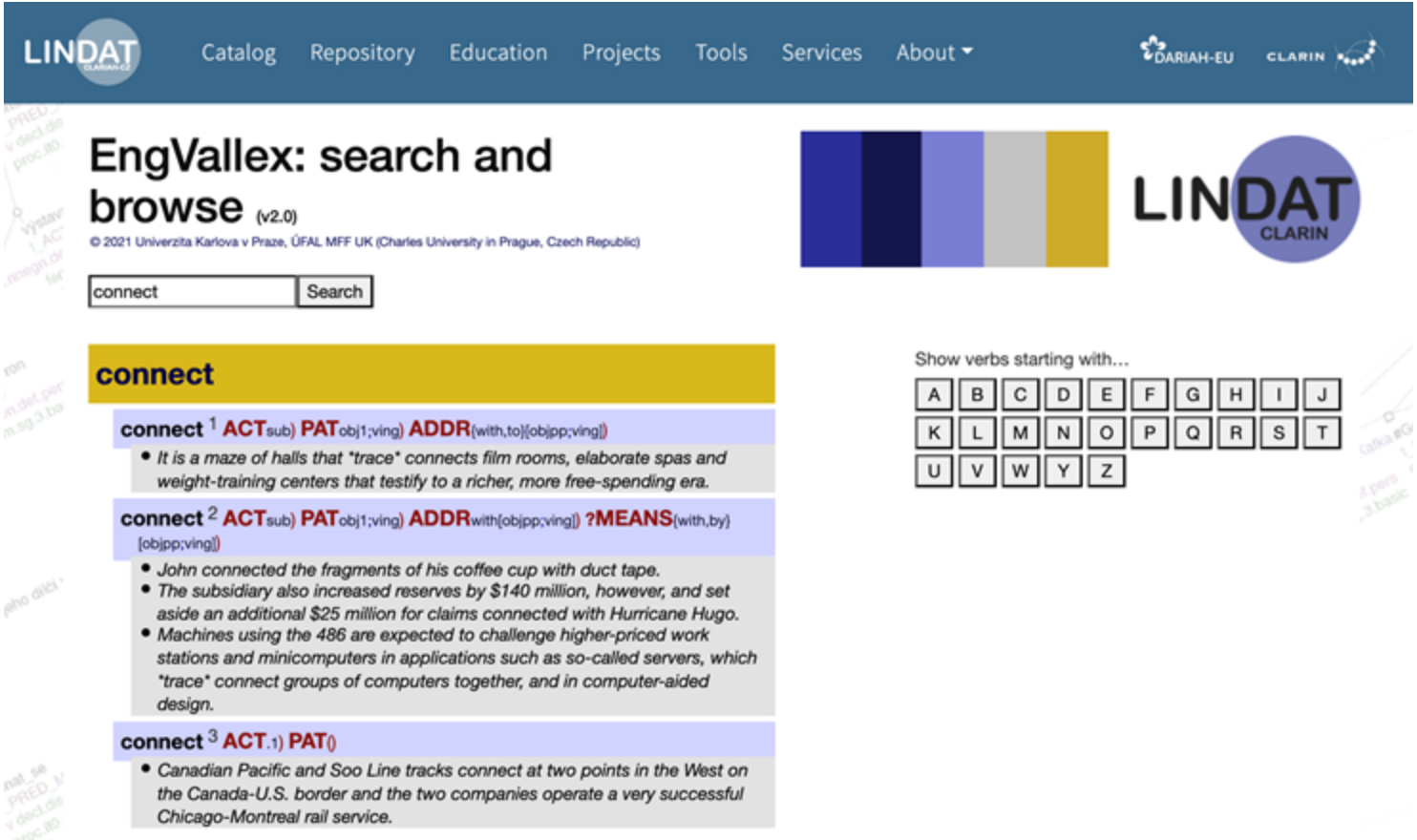
Concordância

EBDL2(776):	Calm and svelte, stealthy as a **cat** in his movements, he seemed to approach sex as a form of research, favouring techniques of foreplay so subtle and prolonged that Robyn occasionally dozed off in the middle of them, and would wake with a guilty start to find him still crouched studiously over her body, fingering it like a box of index cards.	Calmo e ágil, furtivo como um gato nos seus movimentos, parecia abordar o sexo como uma forma de investigação, preferindo técnicas de estimulação tão subtis e prolongadas que Robyn, por vezes, passava pelo sono, e acordava com um sobressalto de culpa para dar com ele ainda inclinado aplicadamente sobre o seu corpo, dedilhando-o como a um ficheiro.
EBDL4(1256):	Angela looked beautiful and Dennis looked like the **cat** who was finally certain of getting the cream.	Angela estava linda e Dennis tinha o ar do gato que sabe finalmente que vão dar-lhe a nata.
EBDL6(2191):	There sit the two men in their familiar attitudes, like **cat** and mouse, spider and fly, the one crouched over his computer console, the other watching from his glass cubicle, his hand moving rhythmically from a bag of potato chips to his mouth and back again.	Lá continuam os dois homens, com os seus típicos comportamentos, como gato e rato, aranha e mosca, um curvado para a consola do seu computador, o outro observando, do seu cubículo de vidro, a mão movendo-se ritmadamente de um pacote de batatas fritas para a boca e vice-versa.
EBJB2(559):	You don't expect a **cat** suddenly to start barking, do you, or a pig to start lowing?	Vocês não esperam, por exemplo, que um gato se ponha de repente a ladrar, ou que um porco desate a mugir, pois não?
EBJB3(875):	Buy a **cat**, own a budgie, but don't hang out with pye-dogs.	Compra-se um gato, um passarinho, mas com cães vadios não se anda mais.
EBJC1(366):	The knitting old woman with the **cat** obtruded herself upon my memory as a most improper person to be sitting at the other end of such an affair.	A velha tricotadora com o gato atravessou-se na minha memória, embora fosse a pessoa menos indicada para me surgir sentada no lado de lá desta história.

Lexica

Lexica contain a **lexical inventory with specific linguistic information** and are primarily used in **NLP applications**. You can find many lexica in the CLARIN Resource Families, out of which, the majority are monolingual. Some of the available lexica offer a browsing interface.

Reference: Cinková, Silvie; Fučíková, Eva; Šindlerová, Jana; et al., 2021, EngVallex - English Valency Lexicon 2.0, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3526.



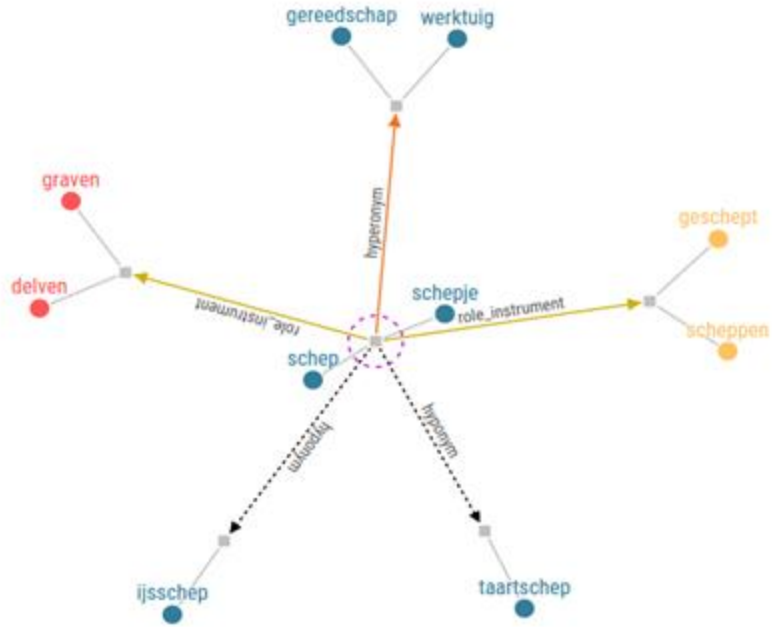
Example of a search in the EngVallex lexicon hosted on [LINDAT/CLARIAH-CZ]

Conceptual Resources

Conceptual resources include lexical resources such as **wordnets**, **thesauri** and **ontologies**. The resources are usually interlinked with semantic relations (e.g. hypernym, hyponymy). Wordnets can be used for word sense disambiguation, machine translation, document clustering.

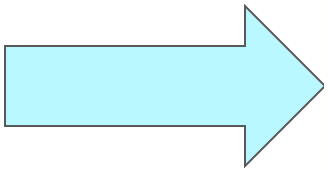
Cornetto is **a lexical resource for the Dutch language** which combines two resources with different semantic organisations: **the Dutch Wordnet with its synset organisation** and the **Dutch Reference Lexicon** which includes definitions, usage constraints, selectional restrictions, syntactic behaviours, illustrative contexts, etc.

https://cornetto.clarin.inl.nl/index.html



How to ensure

- Long term deposit and preservation
- metadata quality
- findability
- citability
- clear licenses
- interoperability



CLARIN

Common Language Resources and Technology Infrastructure

What is CLARIN?

CLARIN is **a distributed digital infrastructure** which provides easy and sustainable **access to a broad range of language data and tools** to support research in the humanities and social sciences and beyond.

CLARIN provides access to multimodal digital language data (text, audio, video) and advanced tools with which to explore, analyse or combine these datasets.

CLARIN Today

- A distributed network of 72 centres (see the Centre Registry)
- 24 members: AT, BE, BG, CY, CZ, DK, EE, ES, FI, GR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, SE, SI, ZA
- 2 observers: CH, UK
- Third Party: Carnegie Mellon University, USA

https://centres.clarin.eu/map

An example of national node

data

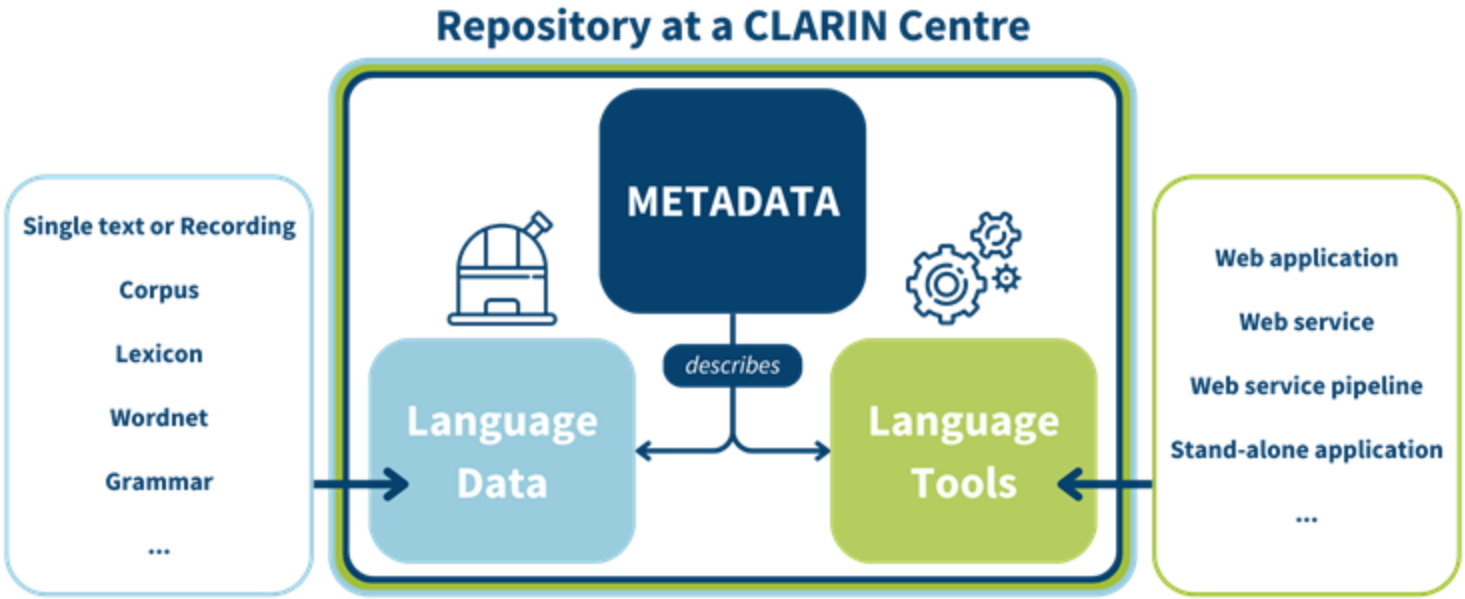
services

knowledge

training

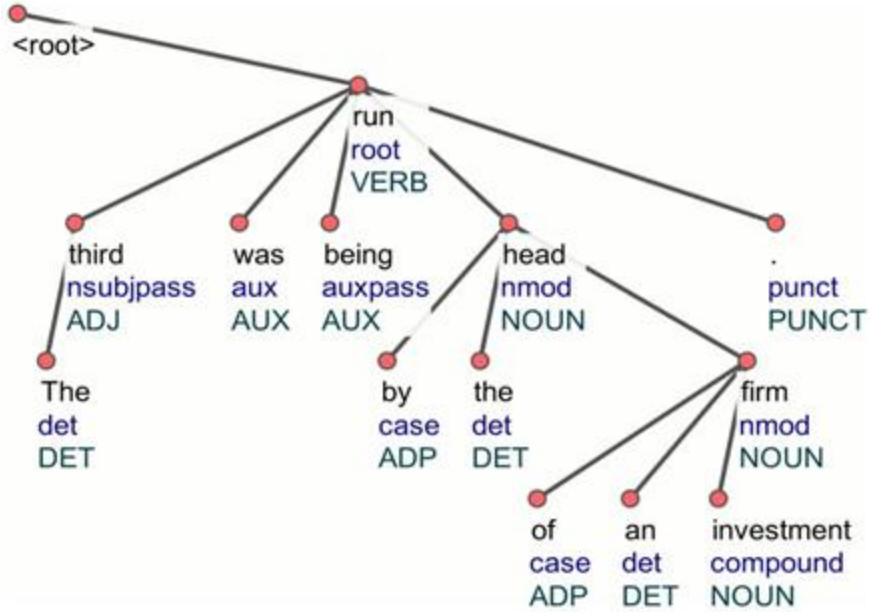
https://portulanclarin.net/

The technical infrastructure



CMDI: Component Metadata Infrastructure https://www.clarin.eu/content/cmdi-component-metadata-infrastructure

Language Resources in CLARIN



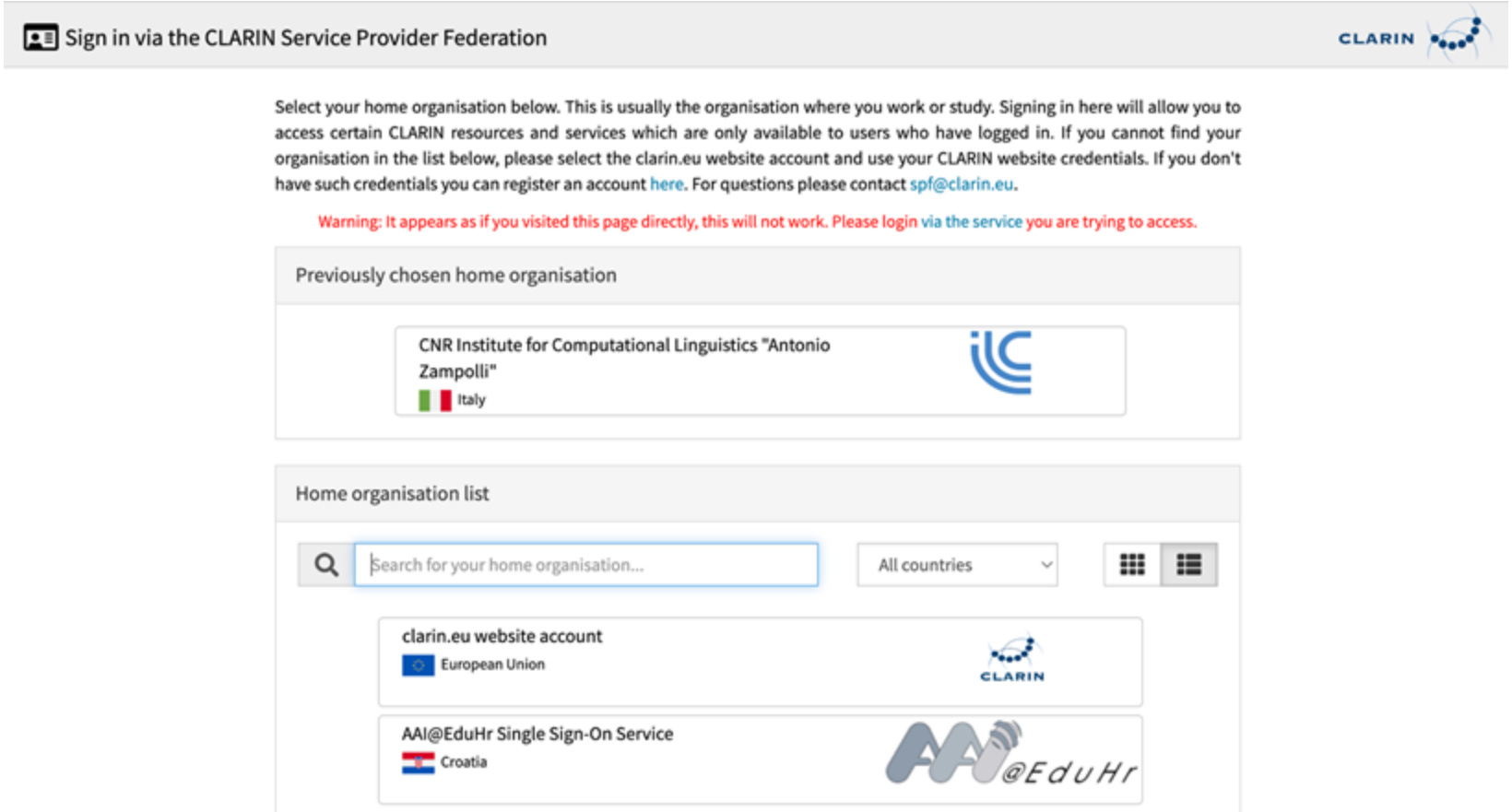
CLARIN core services

- Depositing services to make sure that language resources can be archived and made available to the community in a sustainable way
- The Virtual Language Observatory provides an easy-to-use interface, allowing for a uniform search and discovery process for many resources from a wide variety of domains.
- The Federated Content Search is a search engine that connects to the local data collections that are available in the centres
- The Language Resource Switchboard helps users to find a matching language processing web application for your data
- The Virtual Collection Registry provides a registry where scholars can create and publish their virtual collections

How to access CLARIN services

All users can **freely explore the CLARIN core services** to search for language resources and expertise. Due to license restrictions, some resources are only available for academic users and login is required **using your institutional credentials or CLARIN credentials**

Academic users in all participating countries can access and use the language resources available in CLARIN data centres with a single sign-on access through the CLARIN Service Provider Federation using their institutional credentials


Sign in via the CLARIN Service Provider Federation
CLARIN
Select your home organisation below. This is usually the organisation where you work or study. Signing in here will allow you to access certain CLARIN resources and services which are only available to users who have logged in. If you cannot find your organisation in the list below, please select the clarin.eu website account and use your CLARIN website credentials. If you don't have such credentials you can register an account here. For questions please contact spf@clarin.eu.
Warning: It appears as if you visited this page directly, this will not work. Please login via the service you are trying to access.
Previously chosen home organisation
CNR Institute for Computational Linguistics "Antonio Zampolli"
Italy
Home organisation list
Search for your home organisation...
All countries
clarin.eu website account
European Union
CLARIN
AAI@EduHr Single Sign-On Service
Croatia
AAI@EduHr

Discover and search Language Resources

Search across repositories from all over Europe and beyond with CLARIN ERIC Language Resource Switchboard:

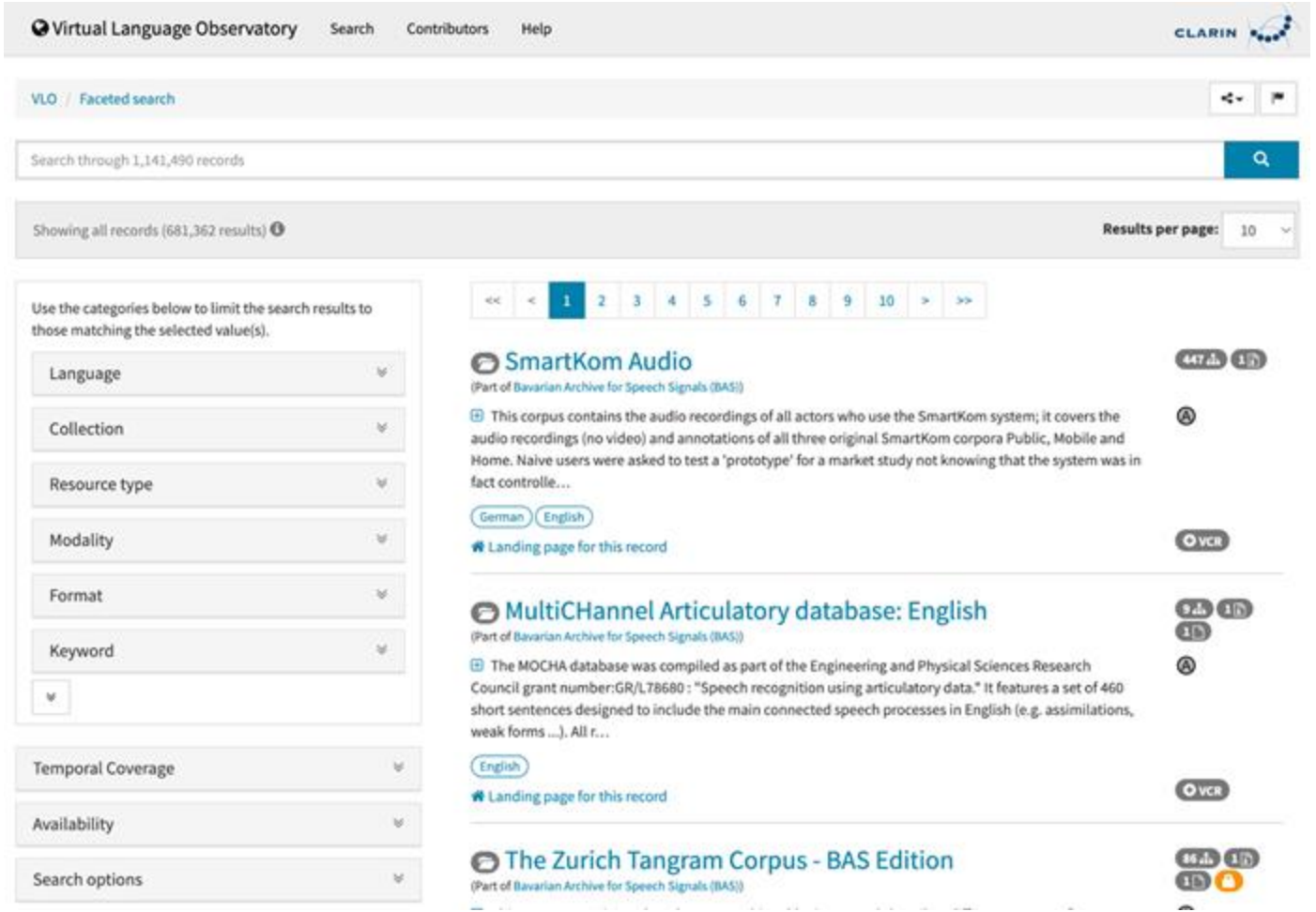
- **A catalogue that harvests metadata** about language resources available in distributed repositories
- **It does not contain language resources**; it just helps you locate it via persistent identifiers
- Even if a resource has restricted access, **the metadata is always freely accessible**
- It uses faceted search to narrow down your searches

https://vlo.clarin.eu/?0

Practice: guided search through the VLO

1) Access the Virtual Language Observatory https://vlo.clarin.eu/?0
2) Refine your search with the faceted options (language, format)
3) Pick a Language Resource and open the record
4) Browse through the metadata and license specification
5) Go to the landing page of the resource
6) Find its recommended citation

Exercise icons created by Freepik - Flaticon


Virtual Language Observatory
Search
Contributors
Help
CLARIN
VLO / Faceted search
Search through 1,141,490 records
Showing all records (681,362 results)
Results per page: 10
Use the categories below to limit the search results to those matching the selected value(s).
Language
Collection
Resource type
Modality
Format
Keyword
Temporal Coverage
Availability
Search options
<< < 1 2 3 4 5 6 7 8 9 10 > >>
SmartKom Audio
(Part of Bavarian Archive for Speech Signals (BAS))
This corpus contains the audio recordings of all actors who use the SmartKom system; it covers the audio recordings (no video) and annotations of all three original SmartKom corpora Public, Mobile and Home. Naive users were asked to test a 'prototype' for a market study not knowing that the system was in fact controlle…
German
English
Landing page for this record
VCR
MultiCHannel Articulatory database: English
(Part of Bavarian Archive for Speech Signals (BAS))
The MOCHA database was compiled as part of the Engineering and Physical Sciences Research Council grant number:GR/L78680 : "Speech recognition using articulatory data." It features a set of 460 short sentences designed to include the main connected speech processes in English (e.g. assimilations, weak forms ...). All r…
English
Landing page for this record
VCR
The Zurich Tangram Corpus - BAS Edition
(Part of Bavarian Archive for Speech Signals (BAS))

COLT - The Bergen Corpus of London Teenage Language (with audio recordings)

Name	COLT - The Bergen Corpus of London Teenage Language (with audio recordings)
Description	COLT is a corpus of London Teenage Language with audio recordings. It is now distributed via the search engine Corpuscle. Corpuscle allows you to pass queries to the corpus, and you may ask for concordances, collocations and distribution. The corpus results from the project COLT. The aim of the project was to create a corpus of British English spontaneous teenage talk and make it available for research, first on the internet, next as an orthographically and prosodically transcribed CD-ROM version, and finally as a CD-ROM version with both text and sound. The recordings were made by 31 volunteering 13-17 year old boys and girls from five socially different school boroughs, so-called 'recruits' equipped with a Sony Walkman, a lapel microphone and a log book. The entire material of roughly half a million words was orthographically transcribed by trained transcribers employed by the Longman Group for transcribing The British National Corpus (BNC). A copy of this version of COLT was incorporated in the BNC. At the Bergen end, the orthographically transcribed material was subsequently submitted to careful editing, which involved correcting misinterpreted talk, reducing the number of <unclear> passages and adding untranscribed talk. The edited version was then tagged for word classes in the same way as the BNC by a research team at Lancaster university.
Collection	CLARINO UiB - Corpuscle
Language	English
Modality	spokenlanguage writtenlanguage
Organisation	University of Bergen CLARINO Bergen Centre

HDL 11495/D9B6-13F8-41BB-1

Landing page

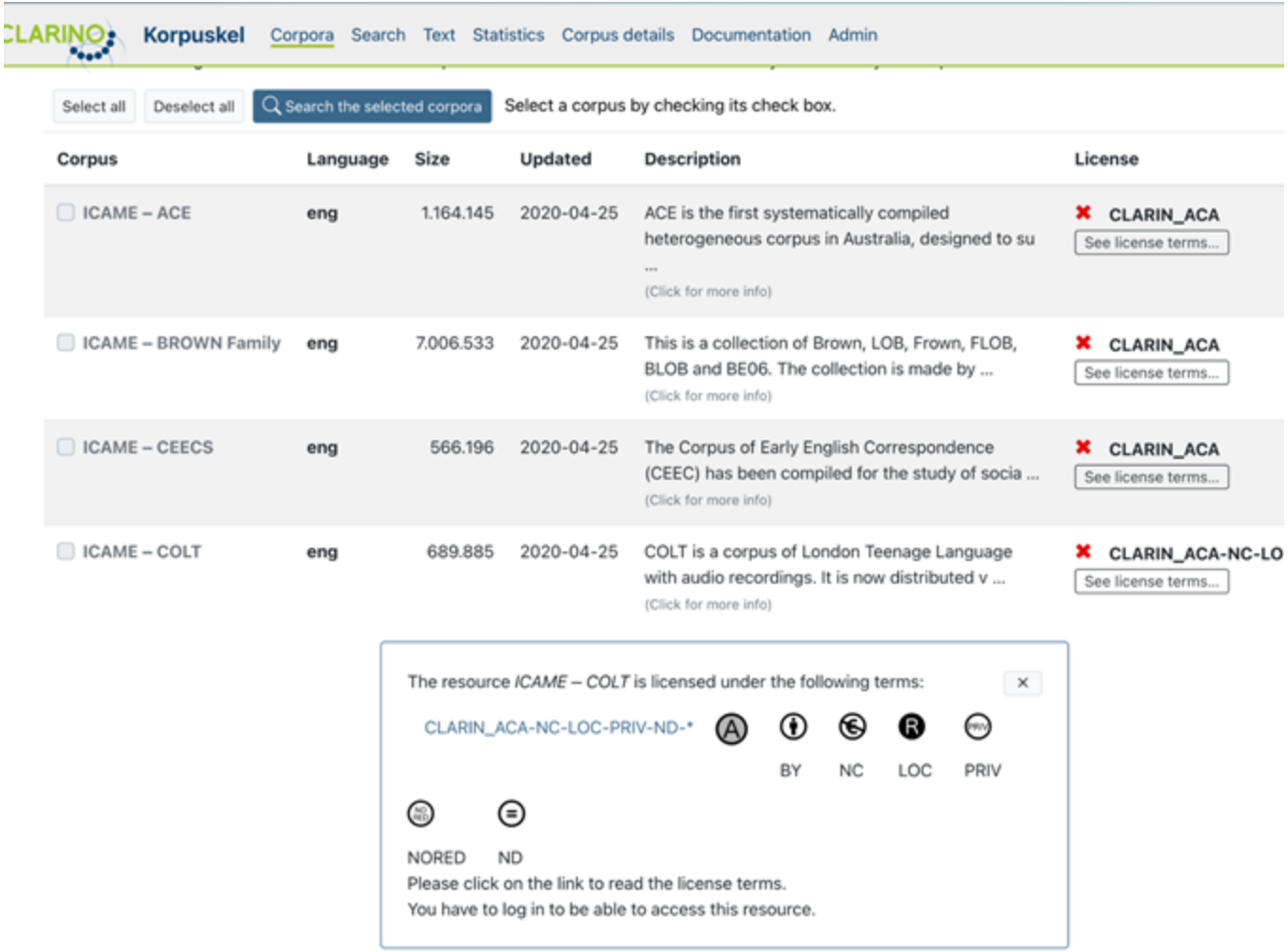
Linked resource

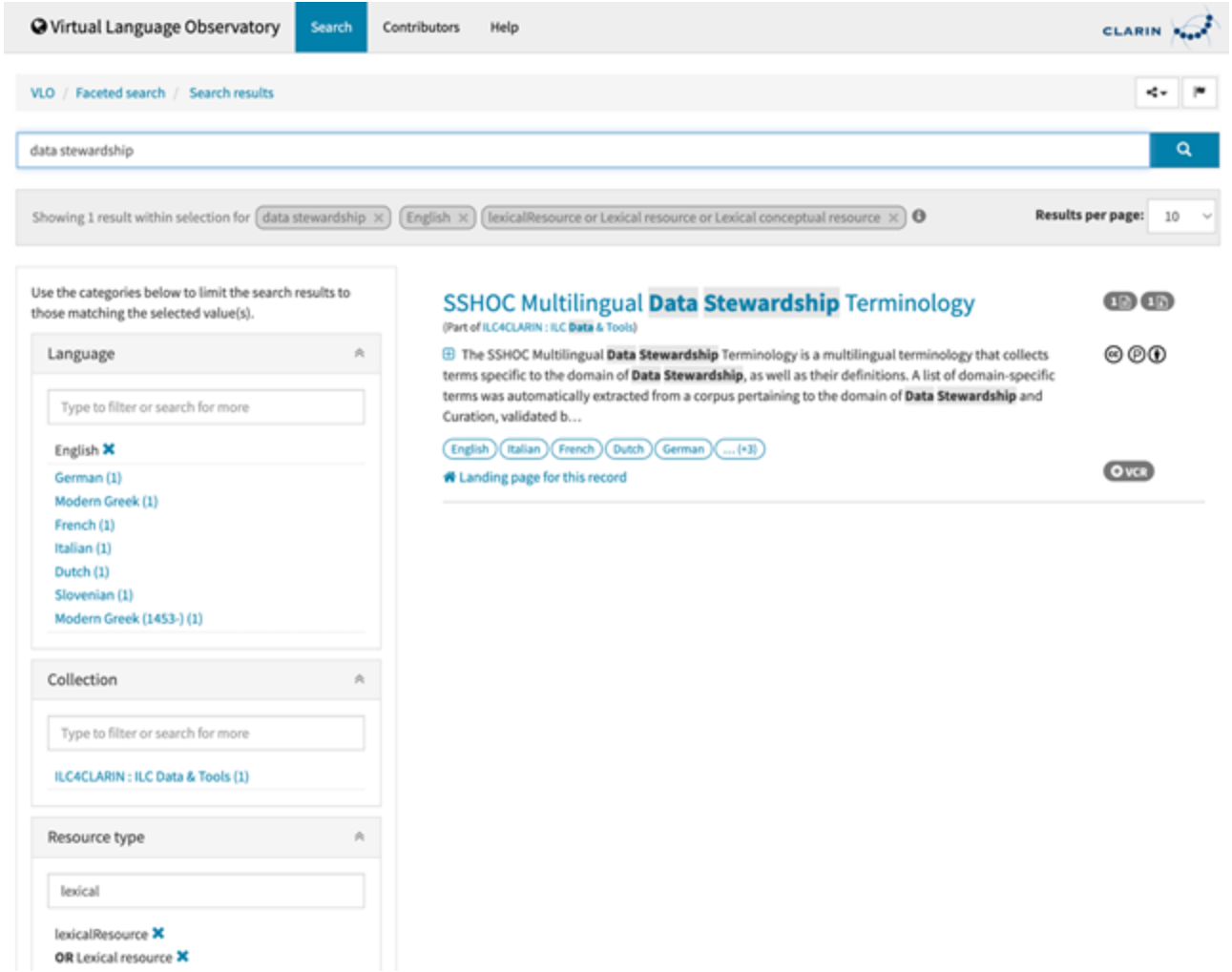
landing-page

HDL 11495/D9B6-13F8-41BB-1

VCR

Add this record to the Virtual Collection submission queue


CLARINO
Korpuskel
Corpora
Search
Text
Statistics
Corpus details
Documentation
Admin
Select all
Deselect all
Search the selected corpora
Select a corpus by checking its check box.
Corpus
Language
Size
Updated
Description
License
ICAME – ACE
eng
1.164.145
2020-04-25
ACE is the first systematically compiled heterogeneous corpus in Australia, designed to su ...
(Click for more info)
CLARIN_ACA
See license terms...
ICAME – BROWN Family
eng
7.006.533
2020-04-25
This is a collection of Brown, LOB, Frown, FLOB, BLOB and BE06. The collection is made by ...
(Click for more info)
CLARIN_ACA
See license terms...
ICAME – CEECS
eng
566.196
2020-04-25
The Corpus of Early English Correspondence (CEEC) has been compiled for the study of socia ...
(Click for more info)
CLARIN_ACA
See license terms...
ICAME – COLT
eng
689.885
2020-04-25
COLT is a corpus of London Teenage Language with audio recordings. It is now distributed v ...
(Click for more info)
CLARIN_ACA-NC-LO
See license terms...
The resource ICAME – COLT is licensed under the following terms:
CLARIN_ACA-NC-LOC-PRIV-ND-*
BY
NC
LOC
PRIV
NORED
ND
Please click on the link to read the license terms.
You have to log in to be able to access this resource.


Virtual Language Observatory
Search
Contributors
Help
CLARIN
VLO / Faceted search / Search results
data stewardship
Showing 1 result within selection for data stewardship English lexicalResource or Lexical resource or Lexical conceptual resource
Results per page: 10
Use the categories below to limit the search results to those matching the selected value(s).
Language
Type to filter or search for more
English
German (1)
Modern Greek (1)
French (1)
Italian (1)
Dutch (1)
Slovenian (1)
Modern Greek (1453-) (1)
Collection
Type to filter or search for more
ILC4CLARIN : ILC Data & Tools (1)
Resource type
lexical
lexicalResource
OR Lexical resource
SSHOC Multilingual Data Stewardship Terminology
(Part of ILC4CLARIN : ILC Data & Tools)
The SSHOC Multilingual Data Stewardship Terminology is a multilingual terminology that collects terms specific to the domain of Data Stewardship, as well as their definitions. A list of domain-specific terms was automatically extracted from a corpus pertaining to the domain of Data Stewardship and Curation, validated b…
English Italian French Dutch German … (+3)
Landing page for this record
VCR

SSHOC Multilingual Data Stewardship Terminology

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

Frontini, Francesca; Gamba, Federica; Monachini, Monica and Broeder, Daan, 2021, *SSHOC Multilingual Data Stewardship Terminology*, ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, http://hdl.handle.net/20.500.11752/ILC-567.

Share:

ILC

Authors	Frontini, Francesca ; Gamba, Federica ; Monachini, Monica and Broeder, Daan
Item identifier	http://hdl.handle.net/20.500.11752/ILC-567
Project URL	https://www.sshopencloud.eu/
Demo URL	https://vocabs.sshopencloud.eu/vocabularies/sshocterm/
Date issued	2021-12-31
Type	lexicalConceptualResource, text
Size	210 concepts
Language(s)	Dutch , English , French , German , Italian , Modern Greek (1453-) , Slovenian
Description	The SSHOC Multilingual Data Stewardship Terminology is a multilingual terminology that collects terms specific to the domain of Data Stewardship, as well as their definitions. A list of domain-specific terms was automatically extracted from a corpus pertaining to the domain of Data Stewardship and Curation, validated by domain experts, assigned a definition, and linked to other existing terminologies (Loterre Open Science Thesaurus, terms4FAIRskills, Linked Open Vocabularies, ISO terms and definitions). Each term-definition pair was then automatically translated into multiple languages (Dutch, French, German, Greek, Italian, Slovenian) by employing Deep-L. The Multilingual Data Stewardship Terminology thus consists of 210 concepts available in Dutch, French, German, Greek, Italian, Slovenian. This resource was created within the frame of the SSHOC (Social Sciences and Humanities Open Cloud) project (H2020-INFRAEOSC-2018-2-823782). It is the result of the work of Task 3.1.2 "extraction of terminology from technical documentation about standards and interoperability", as described in D3.9, carried out jointly by ILC-CNR and CLARIN ERIC.
Publisher	Istituto di Linguistica Computazionale "A. Zampolli" - Consiglio Nazionale delle Ricerche (ILC-CNR)

Practice

Exercise 1

- Go back to ELRC-SHARE repository and the research you did before on how it classifies Language resources
- Go to the VLO and answer the same questions:
  - how does it classify LRs?
  - What types can you find?
  - Take a closer look at the categories: how are they further categorised?
  - How many media are available?
- Compare the two approaches and discuss the differences

Exercise icons created by Freepik - Flaticon

How is it possible?

Discovering Language Resources across Europe and beyond: thanks to the management of data according to the **FAIR Principles**

«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 none 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: https://www.go-fair.org/fair-principles/

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.

CLARIN for FAIR Linguistic Data

In CLARIN, data are:

- **FINDABLE** through access points such as the Virtual Language Observatory and the Federated Content Search
- **ACCESSIBLE** thanks to CLARIN centres providing repositories; through standard metadata sets and a single federated access point
- **INTEROPERABLE** for example in the Language Resource Switchboard, and thanks to the use of controlled vocabularies (Concept Registry) connecting metadata
- **REUSABLE** with the application of open licenses, shared formats and the active curation of metadata

https://www.clarin.eu/fair

FAIR principles and Linked Open Data

In this part, we will see:

1) The Linked Data paradigm and the Linked Open Data Movement
2) The 5-Star Linked Open Data ranking
3) Differences between FAIR and LOD
4) Comparing the 5-star scheme for LOD and the FAIR Principles

«The LD principles emerged as **a series of best practices and principles for exposing, sharing and connecting data on the Web**». Added 2010: «Linked *Open* Data (LOD) is **Linked Data which is released under an open licence**, which does not impede its reuse for free.»

Berners-Lee, Tim. 2006. “Linked Data – Design Issues.” July 27, 2006. https://www.w3.org/DesignIssues/LinkedData.html.

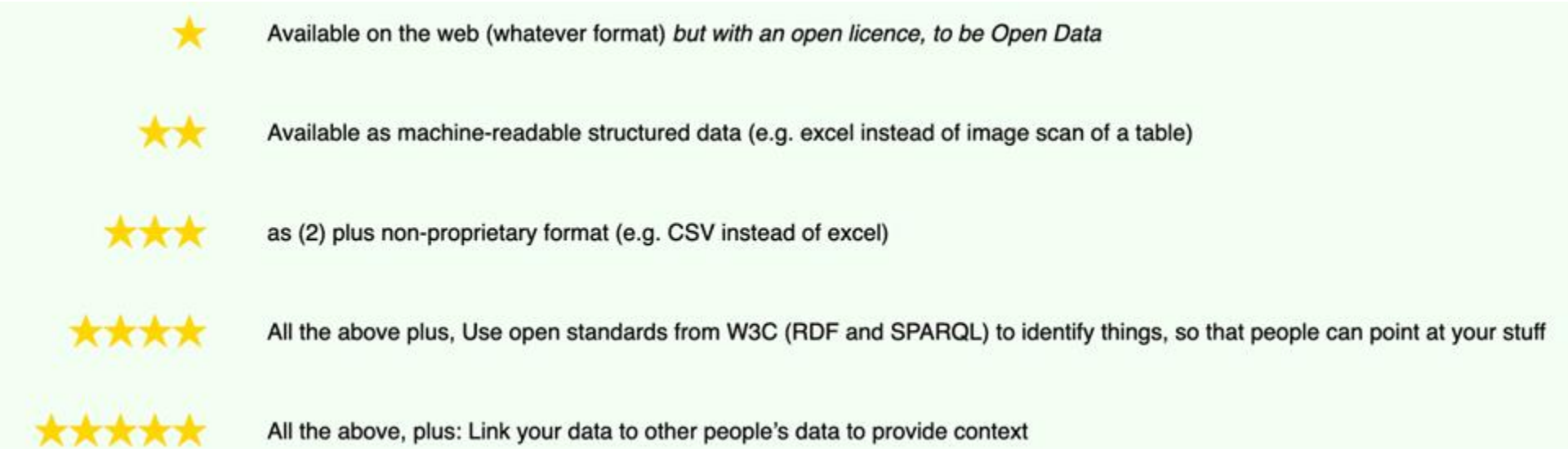
The Linked Data paradigm

The four LD principles are the following:

1) Use URIs as unique names for things
2) Use HTTP URIs so that people and machines can look up those names
3) When someone looks up a URI, provide useful information using Web standards (RDF and SPARQL)
4) Include links to other URIs sot that the user can discover more things

5-Star Linked (Open) Data

«You can have 5-star Linked Data without it being open. However, if it claims to be Linked Open Data then it does have to be open, to get any star at all.» Berners-Lee, added 2010 https://www.w3.org/DesignIssues/LinkedData.html.



FAIR Principles

- Not necessarily open (requires accessibility of metadata)
- Main objective: reusability
- Emphasizes the need for metadata to improve the reusability of data
- Allows for a broader range of PIDs

Linked Open Data

- It mandates open licenses
- Main objective: interoperability
- Metadata are also interoperable data
- Key element: URIs

Source: FAIR Principles and Linked Open Data, Karla Avanço, Te road to FAIR, Hypotheses, 2021 https://roadtofair.hypotheses.org/288

LOD principles can support FAIR principles:

Chiarcos et al., *Linguistic Linked Data*, Springer, 2020, chapter 1.2 Linked Data as an Opportunity to Realize the FAIR Principles, p. 6

- **Findability**: by relying on URIs as globally unique identifiers
- **Accessibility**: by following standard data models such as RDF
- **Interoperability**: by fostering the reuse of existing ontologies and vocabularies, publishing and describing resources in a semantically non-ambiguous ways
- **Reusability**: by adhering to standard data formats and using semantically well-defined vocabularies for describing provenance information, terms of use and licensing conditions

Practice

Exercise 3

Rate the FAIRness of the following resources:

1. https://opensciency.github.io/sprint-content/
2. https://zenodo.org/record/7662732
3. https://www.ebi.ac.uk/training/online/courses/covid-19-data-portal/#vf-tabs__section--overview
4. https://www.markdownguide.org/

Ask yourself: **what info is missing to make the material FAIR?** On what granularity level have you assessed fairness? What type of detailed info is needed so that you can perform a real FAIR assessment?

Thank you for your time and see you tomorrow!

Curly hair icons created by Freepik - Flaticon