

# Introduction to CMC Metadata & Language Data Repositories

Alexander König  
CLARIN ERIC

Workshop on Data Management for FAIR CMC Corpora

27<sup>th</sup> October 2021



# Metadata - Introduction

- Metadata is „Data about Data“
- Should be Structured, Machine-readable
- Helps (your) data to be **F**indable
  - For yourself
  - For others who want to re-use it

# Metadata - Introduction

- Need to find the right balance between
  - Not enough detail
    - Makes it difficult to judge whether a dataset is „what I need“
    - „Fitting“ data might not turn up in a search
  - Too much detail
    - Might overwhelm users
    - Can also make it harder to find data (e.g. data labelled as „Austrian“ might not turn up in a search for „German“ even though the user would probably want it to be found as well)

# Metadata – Standards

- No agreed standard for CMC corpora yet
- Common standards in linguistics
  - Dublin Core/OLAC
  - CMDI
- But there are a lot of others
- Often come with the choice of repository (cf. next section)

# Metadata – Standards - DC

- Dublin Core (DC)
  - Comes from the Library World
  - Very basic information (e.g. author, description, language)
  - Widely supported by software
- OLAC (Open Language Archives Community)
  - Extends DC with additional linguistic MD fields
  - e.g. linguistic field, linguistic data type

# Metadata – Standards - CMDI

- Component MetaData Infrastructure
  - Developed by CLARIN
  - Tries to accommodate the wide variety of users
  - Modular design
    - Metadata profiles
    - Reusable metadata components
- Different CMDI profiles are linked through shared concepts (e.g. „title“ and „name“ will have the same concept (if used to name a resource))

# Metadata – Standards - Findability

- Both OLAC and CMDI come with discovery services as part of their infrastructure
  - OLAC search
  - Virtual Language Observatory (VLO)
- Repositories can register their metadata with one or both (if they can provide it in OLAC/CMDI form)

# Metadata – Standards - Findability

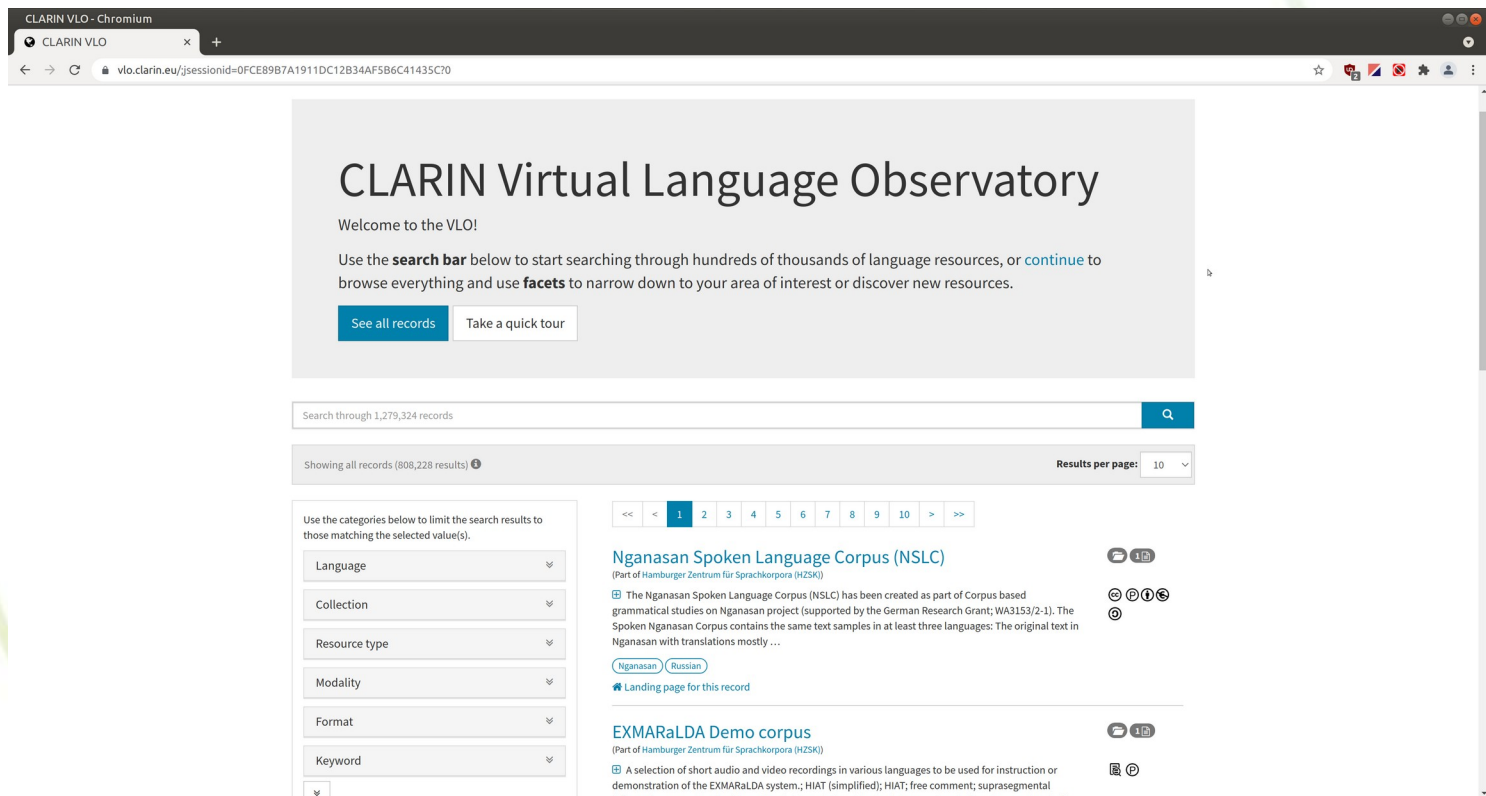
<http://search.language-archives.org/>

The screenshot shows the OLAC Language Resource Catalog website. The page title is "OLAC Language Resource Catalog" and the URL is "dla.library.upenn.edu/dla/olac/index.html". The page features a search bar at the top with the text "Search for language resources" and a "GO" button. Below the search bar, there is a navigation menu on the left with categories like "Navigating the Catalog", "Quick Links", "Contacts", and "More information". The main content area is divided into several sections: "Browse the OLAC records by Geographic region or by Language:" which includes a world map and a list of languages with their respective record counts; "Browse the OLAC records by Archive:" which lists various digital archives and their record counts; and a "Sort by:" section on the right with options for Title, ID, and Date. Below the "Sort by:" section is a "Browse by:" section with multiple expandable categories such as "Archive", "Online", "Subject language", "Language family", "Geographic region", and "Country", each with a list of sub-categories and their record counts.



# Metadata – Standards - Findability

https://vlo.clarin.eu/



The screenshot shows the CLARIN Virtual Language Observatory (VLO) website. The page title is "CLARIN Virtual Language Observatory" and it includes a welcome message: "Welcome to the VLO!". Below this, there is a search bar and instructions: "Use the search bar below to start searching through hundreds of thousands of language resources, or continue to browse everything and use facets to narrow down to your area of interest or discover new resources." There are two buttons: "See all records" and "Take a quick tour".

The search results section shows "Search through 1,279,324 records" and "Showing all records (808,228 results)". The results are displayed in a table with 10 records per page. The first record is "Nnganasan Spoken Language Corpus (NSLC)" (Part of Hamburger Zentrum für Sprachkorpora (HZSK)). The description states: "The Nnganasan Spoken Language Corpus (NSLC) has been created as part of Corpus based grammatical studies on Nnganasan project (supported by the German Research Grant; WA3153/2-1). The Spoken Nnganasan Corpus contains the same text samples in at least three languages: The original text in Nnganasan with translations mostly ...". There are tags for "Nnganasan" and "Russian". A link for "Landing page for this record" is provided.

The second record is "EXMARaLDA Demo corpus" (Part of Hamburger Zentrum für Sprachkorpora (HZSK)). The description states: "A selection of short audio and video recordings in various languages to be used for instruction or demonstration of the EXMARaLDA system.; HIAT (simplified); HIAT; free comment; suprasegmental".

On the left side, there are facets for filtering results by Language, Collection, Resource type, Modality, Format, and Keyword.

# Metadata – Standards - CMC

- No commonly agreed standard
- But apart from general linguistic metadata, some additional fields are „obvious“, most of them pertaining to provenance
  - When was the data collected?
  - Where is the data coming from (e.g. Twitter, Facebook, WhatsApp)?
  - What format is the data in (e.g. full tweets, links to tweets, blog articles)?
  - How was the data processed (anonymized, post-tagged, „cleaned up“ in some way)?
  - What are the licensing conditions?



# Repositories - Introduction

- Research Data Repositories
  - Can be found in most academic institutions
  - Can be general purpose (one repository for the whole university)
  - Or domain-specific (e.g. only hosts linguistic data)

# Repositories – FAIR

- Repositories help with **Findability** and **Accessibility**
- They store data safely (backups) and securely
- Automatically assign a PID
- Force you to add some essential metadata (e.g. license)
- Usually provide easy access, e.g. through single sign-on
- Often propagate information about the data to (domain-specific) search interfaces

# Repositories - Types

- Common repository types
  - Fedora Commons
  - DSpace
  - Dataverse

# Repositories - Instances

- General repositories
  - Zenodo
  - Figshare
- Infrastructures
  - META-SHARE
  - CLARIN Centres

# Repositories - CLARIN

- CLARIN is a network of centres
- Each CLARIN member country hosts at least one centre where you can deposit your data
- Overview:
  - <https://centres.clarin.eu/>
  - <https://www.clarin.eu/content/depositing-services>
- Specialised Centre for CMC data:
  - Eurac Research CLARIN Centre (<https://clarin.eurac.edu/>)



# Repositories - CLARIN

<b>Repository type</b>	<b>Number of centres</b>
DSpace	14
Fedora	10
META-SHARE	4
Git	2
LAT	2
Dataverse	1
Custom	8
<b>TOTAL</b>	<b>41</b>

# Repositories - CLARIN

- All CLARIN Centres provide
  - PID for your corpus
  - Metadata in CMDI and DC (fields are usually fixed)
  - Assistance with depositing (e.g. with writing metadata, choosing a license)
  - Automatic addition to the CLARIN data catalogue (VLO) (and often also to the OLAC search)
  - Authentication and Authorisation through Single Sign-On

# Repositories - Choice

- First have a look at your institute and see if they host a dedicated repository
- If your institution has a general purpose repository, you might use that
- Or look for a domain-specific one outside of your institution
- Talk to your national CLARIN coordinator
- Take a look at the list of CLARIN depositing services
- Ask at the CKCMC or at the ERCC.



Thank you for your attention!

Now it's time for **questions.**