CLARIN

The Language Resources and Technology Infrastructure

Steven Krauwer

CLARIN ERIC / Utrecht University

CLARIN in eight bullets

- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that 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
- that serves as an ecosystem for **knowledge sharing**
- strongly committed to the **FAIR principles**
- and: ready for **integration in EOSC** (European Open Science Cloud; link)

FAIR and EOSC

- The **FAIR principles**:
  - make your data **F**indable
  - make your date **A**ccessible
  - make your data **I**nteroperable
  - make your data **R**e-usable
- The **European Open Science Cloud (EOSC – see next session)** will offer 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States.

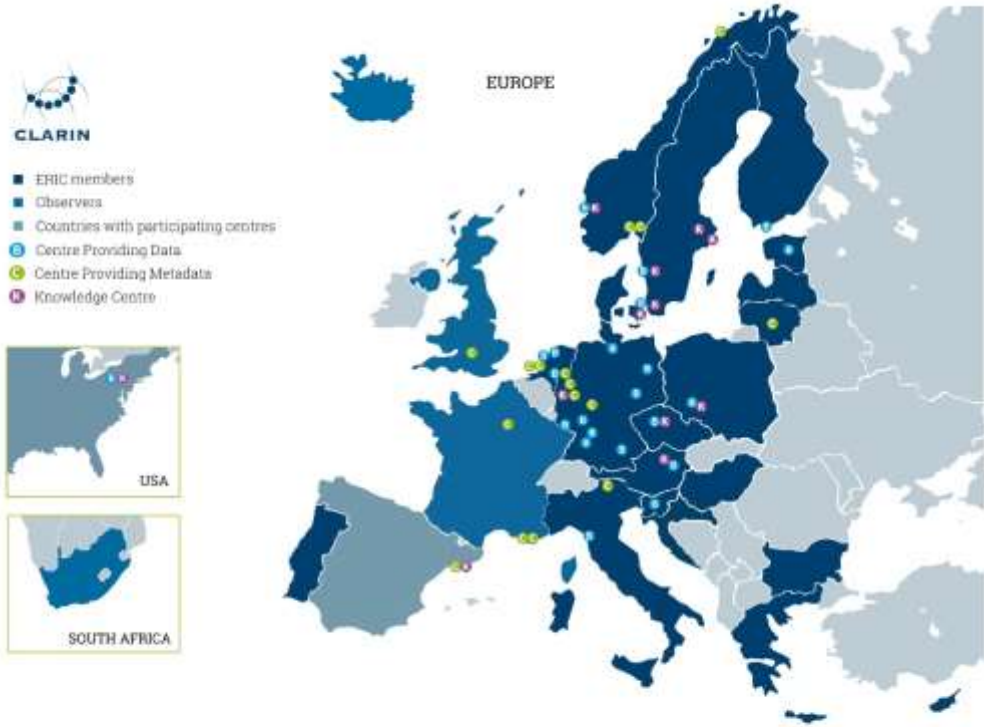
CLARIN ERIC in members and centres

A consortium of:

- 20 member countries:
  AT, BG, CY, CZ, DE,

  DK, EE, FI, GR, HR, HU, IT,

  LT, LV, NL, NO, PL, PT, SE, SI
- 4 observer countries:
  IS, FR, UK, ZA
- >50 centres, subdivided into
  - Centres providing data
  - Centres providing metadata
  - Knowledge centres



CLARIN in services

CLARIN portal

Get an example-based impression of what's currently available

Depositing services

Store language resources in a sustainable repository at a CLARIN centre

Virtual Language Observatory

Discover language resources using a faceted browser or a map

Easy access to protected resources

Get easy access to protected resources, with your institutional username and password.

Language Resource Switchboard

Explore and analyze language data with a wide variety of tools

Virtual Collections

Create your own digital bookmarks, ideal for citing data sets.

Language Resource Inventory

Submit and access information about language resources relevant to your research.

Content Search (prototype)

Search different corpora with a single search engine

Questions & Answers

Searching for a specific data set or application? Wondering how CLARIN can assist your research? Feel free to contact us!

Examples of data types in CLARIN

- Newspaper archives
- Literary texts
- Social Media data
- Parliamentary records
- Historical documents and manuscripts
- Oral History data
- Disciplinary libraries
- Institutional archival data
- Speech and video recordings
- Sign language
- Broadcast archives
- …

See also the info on the CLARIN Resource Families initiative: https://www.clarin.eu/resource-families

Some roles of language ….

- Carrier of cultural content
- Record of the past
- Main communication instrument within and across societies
- Preserving and disseminating our knowledge
- Instrument to formulate rules for society
- Carrier of information
- Means of human expression
- Focus of cognitive processes
- Component of national or cultural identity
- Object of study
- Object of computer processing

Some roles of language …
… and the corresponding audiences

- Carrier of cultural content: *cultural heritage*
- Record of the past: *archaeology, history*
- Main communication instrument within and across societies: *sociology, anthropology*
- Preserving and disseminating our knowledge: *all disciplines*
- Instrument to formulate rules for society: *law, theology*
- Carrier of information: *media studies, journalism*
- Means of human expression: *literary studies, art, psychology*
- Focus of cognitive processes: *brain studies, psychology, neurology*
- Component of national or cultural identity: *political sciences, social sciences*
- Object of study: *linguistics, language learning*
- Object of computer processing: *language and speech technology*

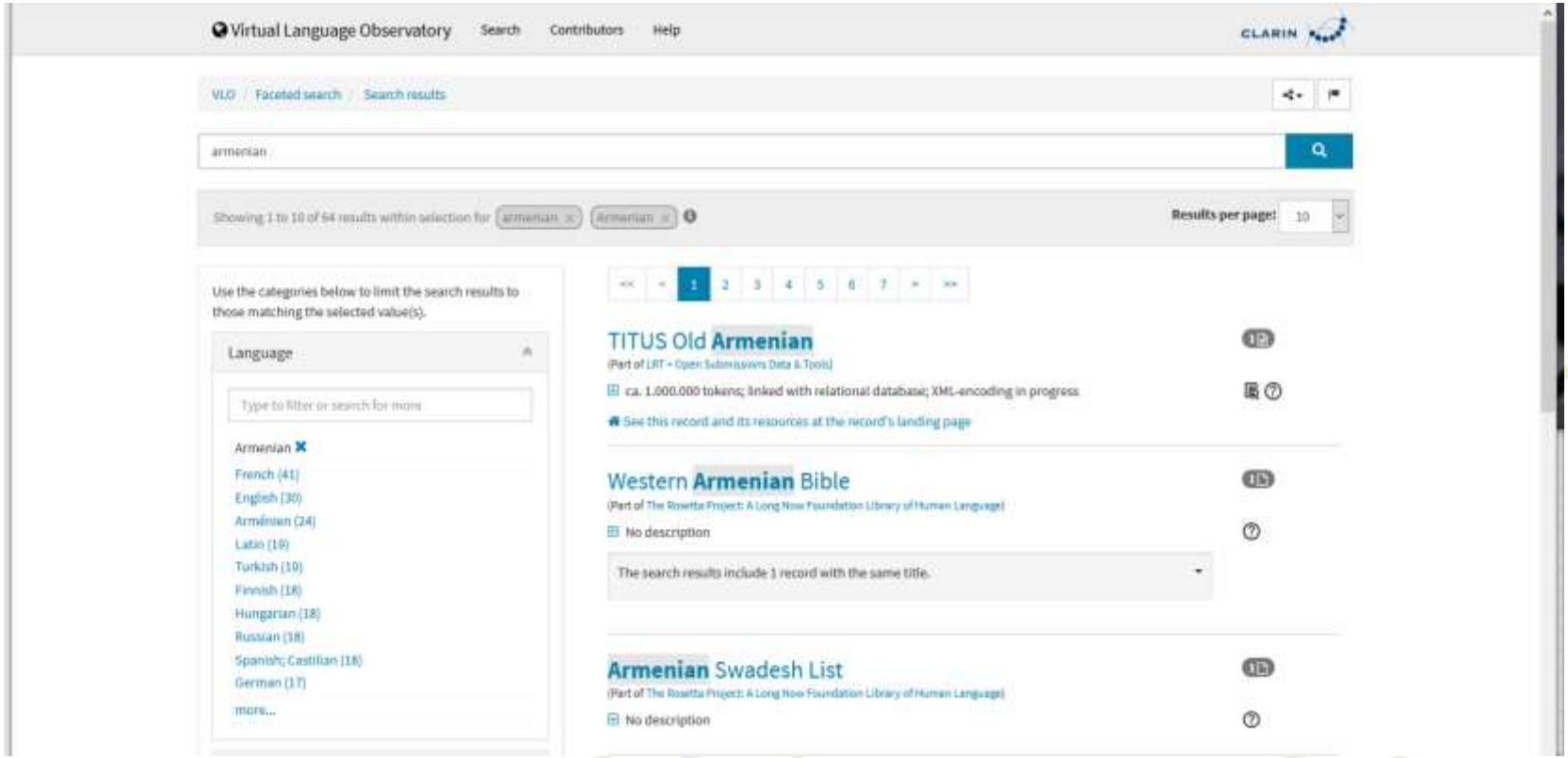
Some roles of language …
… and how these connect with culture

- **Carrier of cultural content: *cultural heritage***
- **Record of the past: *archaeology, history***
- Main communication instrument within and across societies: *sociology, anthropology*
- Preserving and disseminating our knowledge: *all disciplines*
- Instrument to formulate rules for society: *law, theology*
- **Carrier of information and entertainment: movies, *media, journalism***
- **Means of human expression: *literary studies, art, psychology***
- Focus of cognitive processes: *brain studies, psychology, neurology*
- **Component of national or cultural identity: *political sciences, social sciences***
- Object of study: *linguistics, language learning*
- Object of computer processing: *language and speech technology*

Why am I here?

- Repeat from first slide:
  - Infrastructure … that 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*
- The Matenadaran provides an excellent example:
  - It is the world's largest repository of Armenian manuscripts
  - Part of it has been, or is being digitized
  - Armenians and scholars interested in Armenian language and culture are spread all over the world
  - Wouldn't it be great if this growing digital collection could be made widely accessiblefor all scholars, together with tools to work with it?
- CLARIN would be happy to collaborate to make this happen

Only 64 Armenian items in the CLARIN catalogue right now



More generally about collaboration

- CLARIN is very keen on **establishing collaboration** links with repositories sitting on valuable language data in Eastern Europe
- We also want to facilitate **cross border collaboration in research and education**.
- One of our instruments are the **CLARIN Resource Families**: collections of similar data that are likely to exist in all languages, and that can be used to do collaborative, comparative or contrastive research and training across languages, countries and disciplines
- This would also allow for **porting methods and tools between languages**

Examples of Resource Families on the next slide

CLARIN Resource Families

- Some examples:
  - Computer-mediated communication corpora
  - Historical corpora
  - L2 learner corpora
  - Literary corpora
  - Newspaper corpora
  - Parallel corpora
  - Manually annotated corpora
  - Parliamentary corpora
  - Spoken corpora
- More on https://www.clarin.eu/resource-families

To conclude

- CLARIN is an infrastructure focused on language in all its manifestations.
- It serves a broad range of audiences, belonging to different disciplines.
- We want to give access to all relevant data and tools, including yours.
- We invite interested parties in CEE and other countries to collaborate with us.

- Contact:
- www.clarin.eu
- Steven Krauwer (steven@clarin.eu)