

CLARIN, Open Science and industry: What common interests?

Cristina Grisot, CLARIN-CH National Coordinator
Alexandru Craevschi, CLARIN-CH Technical Officer
CLARIN-CH c/o University of Zurich
cristina.grisot@uzh.ch, alexandru.craevschi@uzh.ch



Poster presented at the Swiss NLP Expo, June 13, 2025

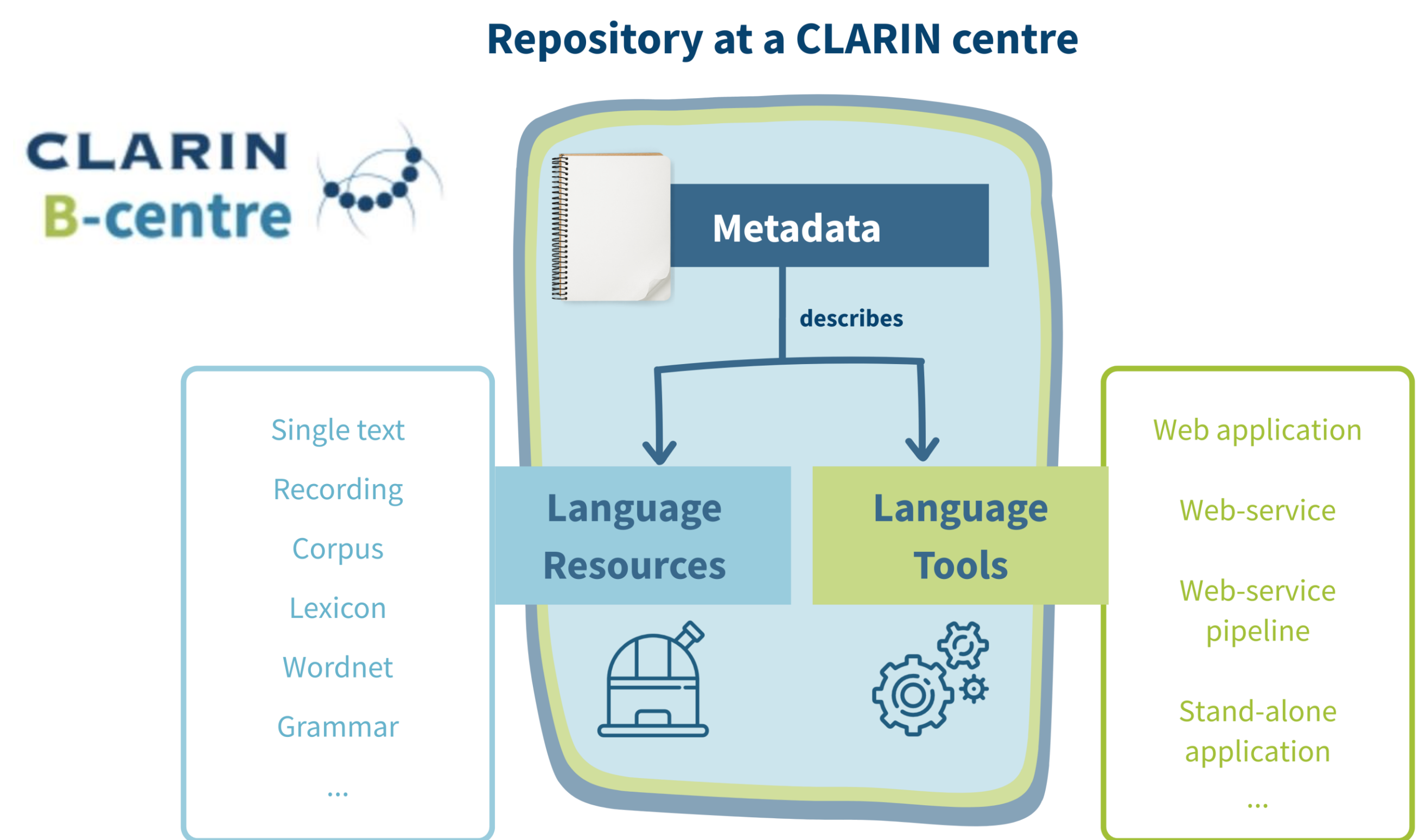
CLARIN: Common Language Resources and Language Technology

- ❖ ESFRI roadmap (2006), a consortium of type ERIC (2012), Landmark (2016)
- ❖ After 12 years: 26 european countries and South Africa are members
- ❖ Provides **easy and sustainable access** for scholars working with language data to
 - ✓ **Diverse digital language data** (in written, spoken, video or multimodal form)
 - ✓ **Tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
 - ✓ A **single sign-on environment** using the CLARIN Federated Identity Provider
 - ✓ Serves as an **ecosystem for knowledge exchange**
- ❖ A **distributed network** of >70 centres, 22 CTS certified data centres, and strong focus on **FAIRness & interoperability**
- ❖ CLARIN is a **fully operational infrastructure for sourcing, processing, and licensing language data at scale**.

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
 - ✓ **Interoperability** through a common metadata framework: CLARIN CMDI
- ❖ Promotion of **responsible data science**
- ❖ Support for **linguistic diversity**
 - ✓ Data covering over 1500 languages
 - ✓ Tools for many languages
 - ✓ Language resources in all modalities
- ❖ Strengthening the support for > 500,000 professional SSH researchers in Europe and globally a multitude

How CLARIN works



- ✓ CLARIN Standards Information System (SIS)
- ✓ CLARIN License Category Calculator

Language resource: ParlaMint

- ❖ Parliamentary proceedings of **29 European parliaments**
- ❖ Subcorpora:
 - ✓ **REFERENCE**: until 30.01.2020
 - ✓ **COVID**: from 31.01.2020 onwards
 - ✓ **WAR**: from 24.02.2022 onwards
- ❖ Linguistically annotated, named entities, rich metadata
- ❖ The ParlaMint corpora are **interoperable**: using specifically developed TEI-based schema
- ❖ The ParlaMint corpora are **comparable**: same periods (2015–2022), same metadata (speakers, parties, sessions, reactions, etc.), same types of linguistic annotations

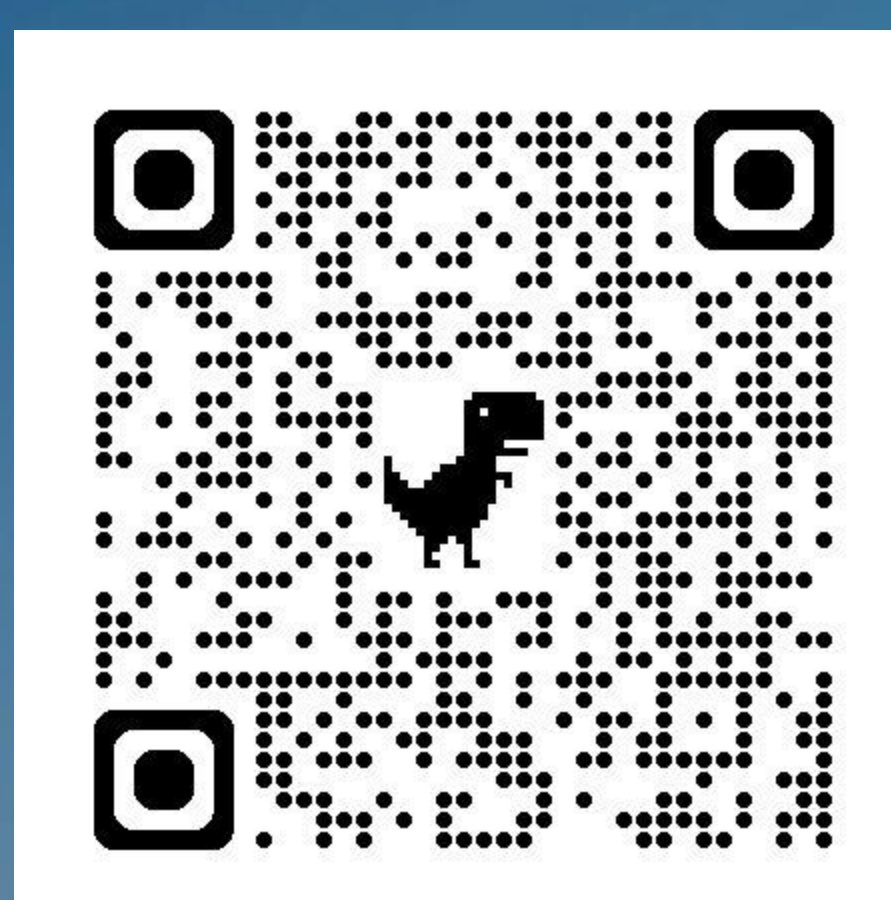
CLARIN and Industry

- ❖ **CLARIN's strategy:** “Develop systematic engagement with **non-academic communities** that have an interest in CLARIN's outputs and can fuel technical or societal innovations”
- ❖ **Findable & Accessible Resources:** CLARIN's VLO indexes tools and datasets from Europe with metadata and unified search
- ❖ Domain, ethical, legal **expertise**, good practices, cross-country committees to establish standards
- ❖ CLARIN promotes **dialogue, joint projects, and hackathons**. R&D services can tap into academic grants, data-science expertise, and student talent pools
- ❖ CLARIN's strength lies in its **pan-European federation**. Each consortium brings unique resources, expertise, and services



CLARIN ERIC was established in 2012 and received ESFRI Landmark status in 2016

www.clarin.eu



 clarin@clarin.eu

 github.com/clarin-eric

Acronyms