CLARIN's Contribution towards an Open Science Infrastructure for the Digital Humanities

Darja Fišer
CLARIN ERIC Director of User Involvement

Open Access in the Humanities
Ljubljana, 22 May 2018

CLARIN in seven bullets

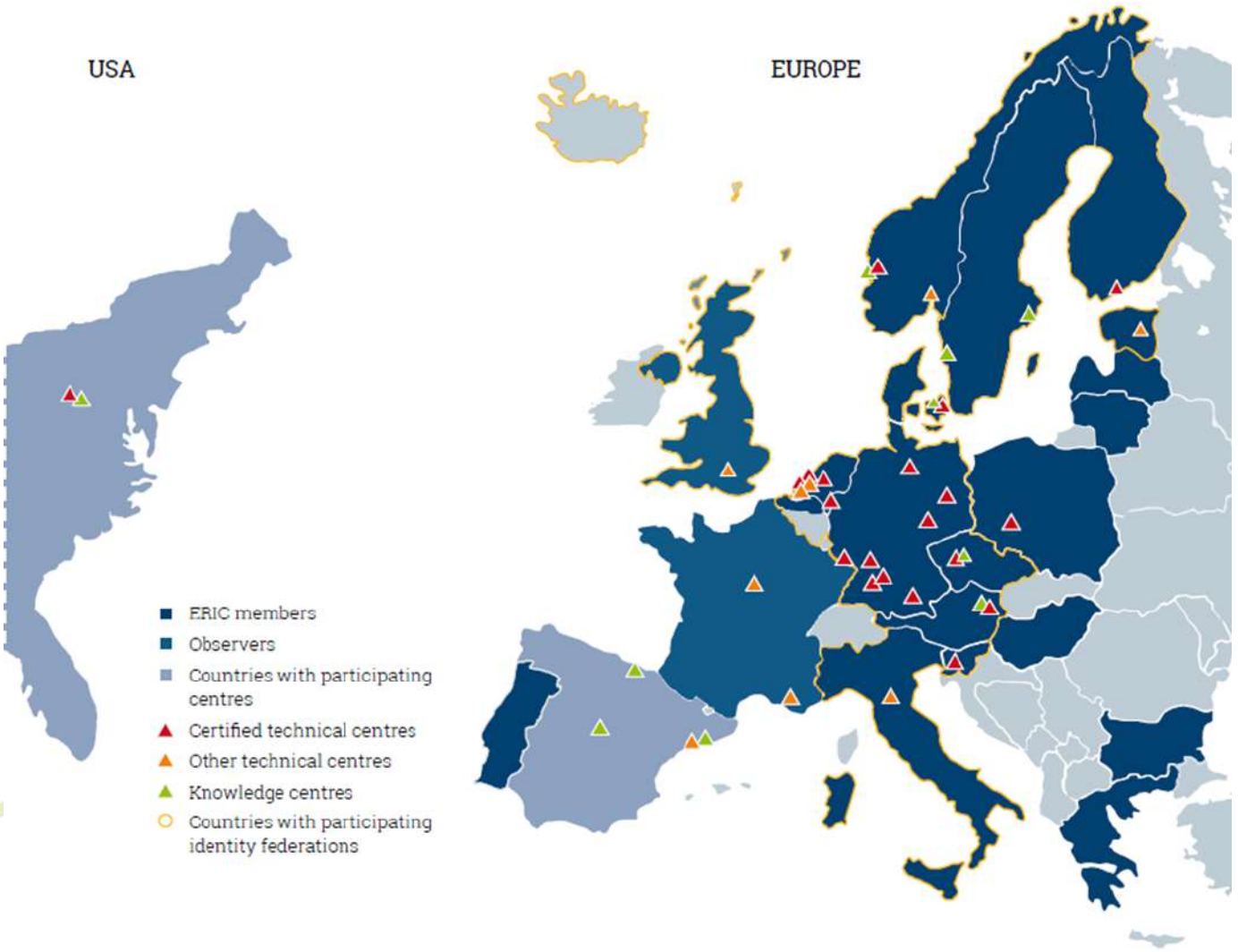
1. **CLARIN** is the Common Language Resources and Technology Infrastructure
2. **ESFRI** ERIC status since 2012, Landmark since 2016
3. that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
4. to **digital language data** (in written, spoken, video or multimodal form)
5. and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
6. through a **single sign-on** environment
7. and that serves as an ecosystem for **knowledge sharing**.

CLARIN ERIC in members and centres

A consortium of:

1. 20 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI
2. 2 observers:

   FR, UK;
3. >40 centres



CLARIN.SI

- **CLARIN.SI member of the CLARIN ERIC General Assembly**: Albin Kralj, Ministry for Education, Science and Sport of the Republic of Slovenia
- **CLARIN.SI National Coordinator**: Tomaž Erjavec, Department of knowledge technologies, Jožef Stefan Institute
- **CLARIN.SI Repository:** 90 resources
- **CLARIN.SI Tools and Services**: concordancers, annotation tools, text processing tools

- **CLARIN.SI Partners**:
  - Alpineon
  - Amebis
  - Domestic Research Society
  - Jožef Stefan Institute
  - Institute of Contemporary History
  - Slovenian Language Technologies Society
  - Trojina
  - University of Ljubljana
  - University of Maribor
  - University of Nova Gorica
  - University of Primorska
  - Scientific and Research Centre of the Slovenian Academy of Sciences and Arts

CLARIN's goal

Digital language resources and tools from all over Europe and beyond are accessible and interoperable for the support of researchers in the humanities and social sciences

CLARIN'S priorities

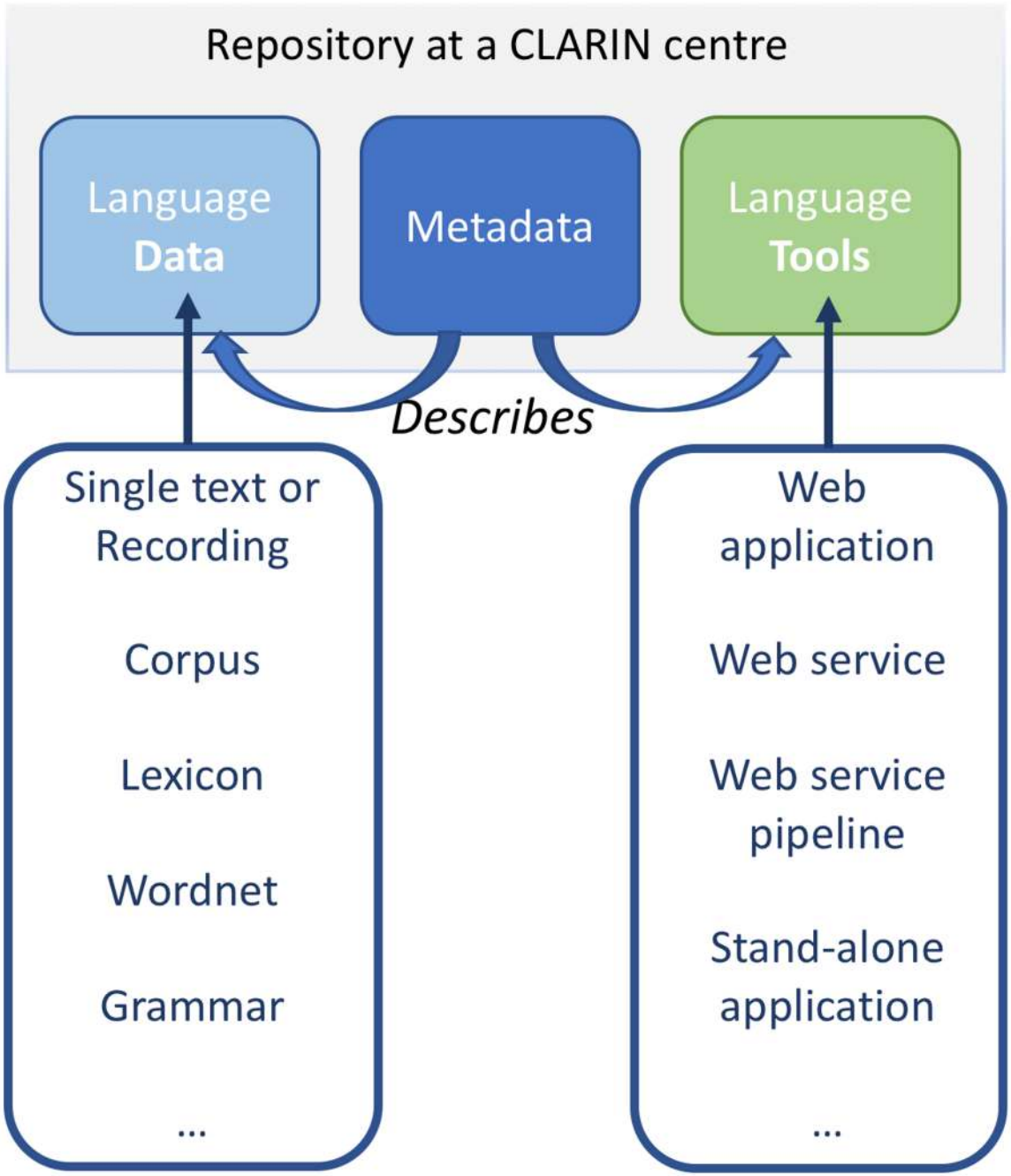
1. **Uptake by researchers**: outreach to all humanities disciplines (researcher training courses, workshops, etc), service enhancements for consistent user experience
2. **Technical infrastructure**: towards an integrated, interoperable infrastructure for Open Science (technical centres, services, licenses etc.)
3. **Knowledge sharing**: knowledge centres, mobility grants, video lectures, course registry (with DARIAH), annual conference
4. **Sustainability**: extension to new countries, cooperation with GLAM sector, commitments from stakeholders and funders, cooperation with other infrastructures

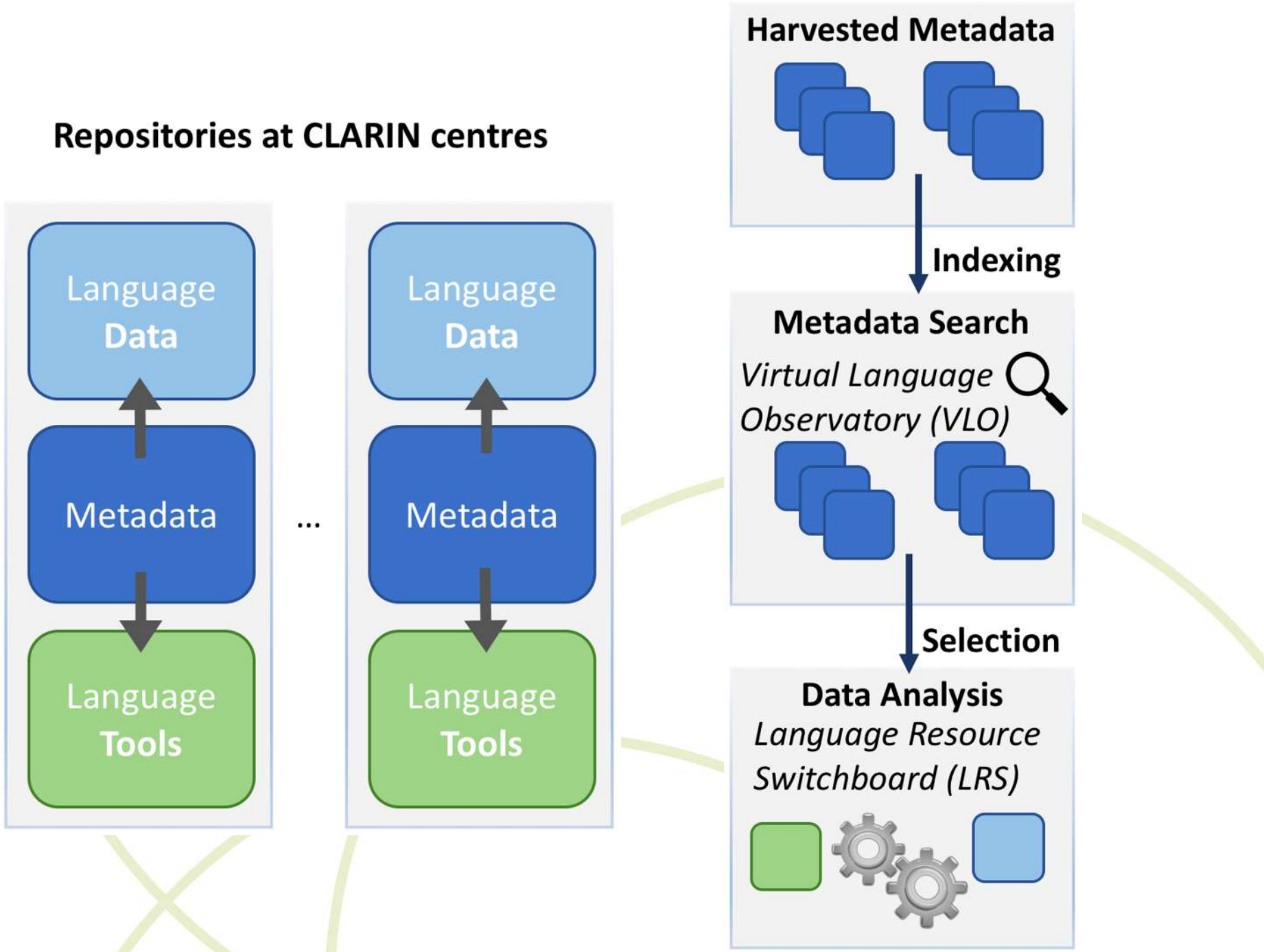
CLARIN for Open Science

Not a stand-alone facility, but a player in making the vision that is underlying the emerging European policies towards Open Science a reality, interconnecting researchers across national and discipline borders by offering seamless access to data and services in line with the FAIR data principles

FAIR principles

- **Findable***:* data must be registered with a persistent ID and items must be collected in a catalogue
  - *CMDI metadata to describe the data and tools in repositories*
  - *these metadata are open and are harvested from centres into catalogues*
- **Accessible***:* open access protocol (subject to restrictions), clear procedure for authentication and authorization
  - *HTTP protocol and SAML for federated single sign-on*
  - *resources are as open as possible, when restrictions are necessary the conditions for use are made explicit*
- **Interoperable***:* documented descriptive vocabulary, standards for data and metadata coding
  - *CMDI framework as common metadata language, links to standardized OpenSKOS vocabularies, standardized linking to datasets and landing pages*
  - *recommendations for the use of standard data formats, such as TEI*
- **Re-usable***:* clear licenses, understandable documentation (including provenance), compatibility with community standards and tools
  - *clear recommendations on license disclosing and user-friendly ways of categorizing these, the provenance needs to be part of the metadata*
  - *while community standards are hard to define, the bottom-up structure of the centres brings along close ties with best practices*


Repository at a CLARIN centre
Language Data
Metadata
Language Tools
Describes
Single text or Recording
Corpus
Lexicon
Wordnet
Grammar
...
Web application
Web service
Web service pipeline
Stand-alone application
...


Repositories at CLARIN centres
Language Data
Metadata
Language Tools
...
Language Data
Metadata
Language Tools
Harvested Metadata
Indexing
Metadata Search
Virtual Language Observatory (VLO)
Selection
Data Analysis
Language Resource Switchboard (LRS)


CLARIN
http://clarin.eu