CLARIN's first steps on the long path to software sustainability

Dieter Van Uytvanck

Technical Director CLARIN ERIC

dieter@clarin.eu

EURISE workshop

Utrecht

12 March 2019

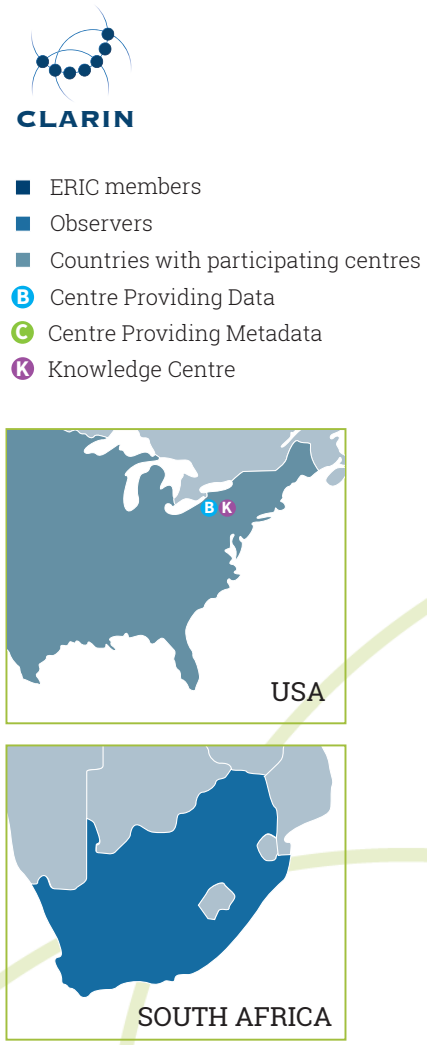
CLARIN in six bullets

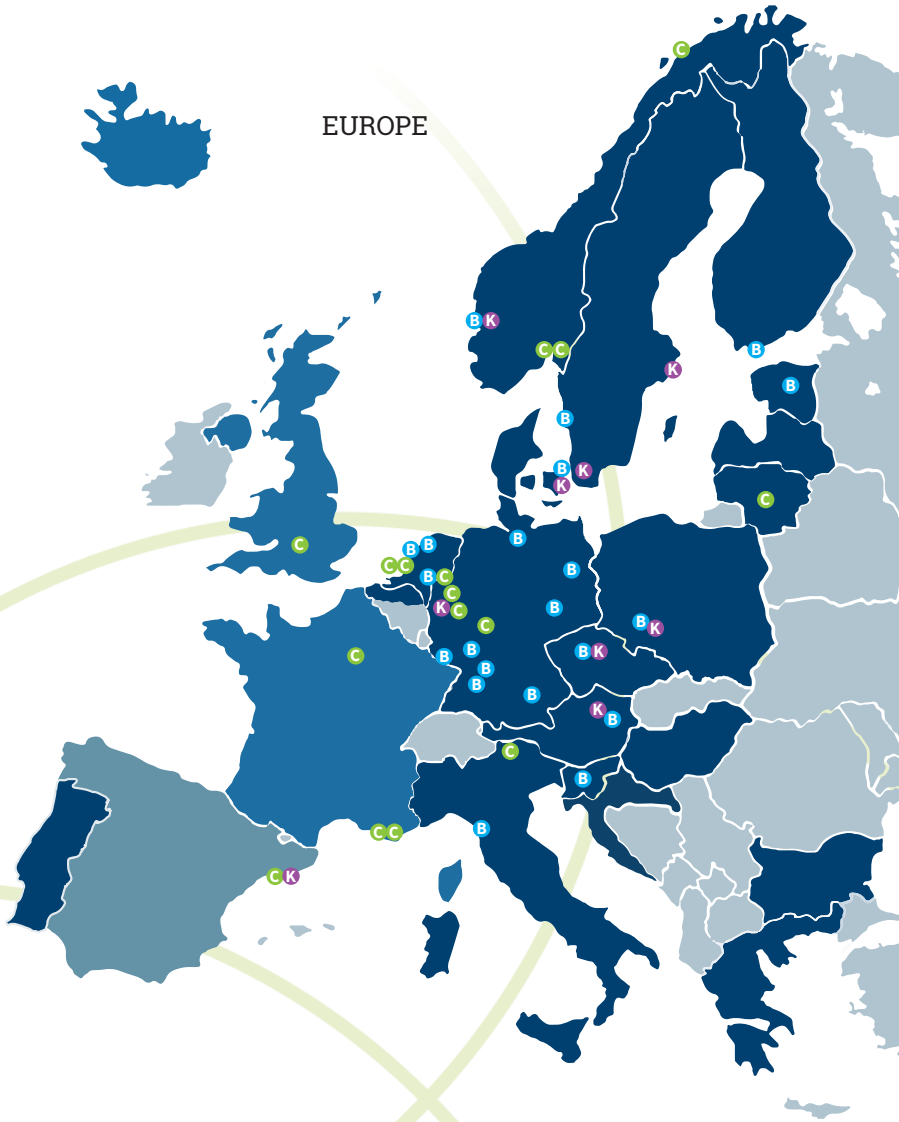
- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
  - to **digital language data**
    - in written, spoken, video or multimodal form
  - to **advanced tools**
    - to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** online environment

CLARIN ERIC in members and centres

A consortium of:

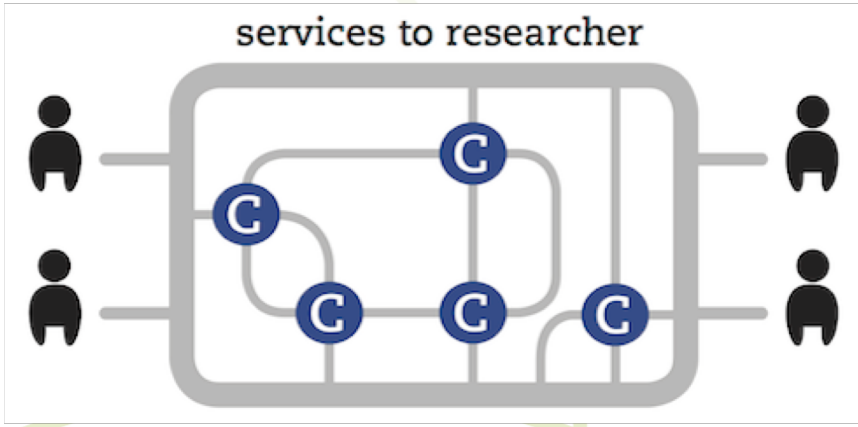
- 20 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI
- 4 observers: FR, UK, IS, SA
- >50 centres





CLARIN centres

- A **distributed architecture**: (http-accessible) files, web applications and web services spread all over Europe
- Nodes in the network: **centres**
- Currently:
  - > 50 registered centres
  - 22 certified B-centres



Technical pillars

- **Federated identity** - letting users login to protected data and services with their own institutional login and password
- **Persistent identifiers** - enabling sustainable citations of electronic resources
- **Sustainable repositories** - digital archives where language resources can be stored, accessed and shared
- **Flexible metadata and concept definitions** - to ensure semantic interoperability when describing language resources
- **Content search** - offering a search engine for a wide range of language resources
- **Web service chaining** - giving users the possibility to freely combine language processing services

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!

Our central developers team

Willem Elbers

André Moreira

Twan Goosen

Menzo Windhouwer

Which software?

- Central services, eg:
  - Virtual Language Observatory – https://vlo.clarin.eu
  - Virtual Collection Registry – https://collections.clarin.eu
  - Language Resource Switchboard – https://switchboard.clarin.eu
  - Federated login discovery service – https://discovery.clarin.eu
    - Compatibility & speed
  - Centre registry – https://centres.clarin.eu
- Distributed services
  - Metadata curation module – https://curate.acdh.oeaw.ac.at
  - Several validators (OAI-PMH, Federated Content Search)
- Centre-specific services and applications > majority
  - Long-term supported (ELAN, WebLicht, UD Pipe, …)
  - Software as a project outcome (sometimes short-lived)

Which issues?

- Lacking awareness
  - “I just needed to reboot the server. I did not know someone was using this service.”
  - “Right. That project ended, and now we have a new tool. It’s much better, btw”
- Living with and surviving hypes and trends
  - SOAP? Flash/Flex?
- Simplifications on the side of the management, leading to a lack of resources
  - “Software Maintenance = server administration”
  - “More Person Months = better software”
  - “It’s all working = the code is OK”
- Organisational factors
  - Developers leaving, including know-how
  - Time pressure: hacks, bad/no documentation

Which solutions?

- **No golden bullet**
- Minimalism:
  - avoid development of generic software/modules
  - difficult!
- Free Software approach
  - Publish soon, often, document, build a community (atmosphere!)
- Redundant know-how
- Standardize deployment and dependencies
  - Maven: make it easy to build
  - Docker + Docker Compose: make it easy to deploy/reproduce
- Testing:
  - unit testing + CI
  - monitoring (eg OAI checks in centre registry)
- On the horizon:
  - Early code peer-review and feedback

Guidelines for software development in projects (1)

- Documentation
- Implementation decisions:
  - Java / Python / Bash
  - Apache Wicket for Java-based web-applications
  - JAX-RS/Jersey for REST services
  - React for Javascript based front-ends
  - Bootstrap for web page/front-end styling (HTML & CSS)
    - Use the Bootstrap based CLARIN base style to match the common CLARIN styling
- Quality and integrity:
  - Code layout/naming/… conventions
  - Automated tests
  - Code review
  - Secrets in environment variables, not in the code

Guidelines for software development in projects (2)

- Organisation:
  - Make yourself known
  - Ask, ask, ask!
- Portability (next slide)

Our Docker approach

- In CLARIN ERIC, as central deployment workflow
  - https://www.clarin.eu/event/2018/centre-meeting
  - https://gitlab.com/CLARIN-ERIC/build-workflow-documentation

- Interestingly enough, this also helped for setting up a framework for reproduction of scientitic data processing workflows:
  - http://wordpress.let.vupr.nl/lrec-reproduction
  - https://gitlab.com/CLARIN-ERIC/reprolang

Conclusions

- Nothing unexpected, nor for the problems, nor for the solutions.
- But that provides a good case for information & best practice sharing in a forum like EURISE!

Thank you for your attention!

- Questions?

- Interested? Go to www.clarin.eu/contact and subscribe to the CLARIN newsflash