CLARIN
Common Language Resources and Technology Infrastructure

DARIAH-EU

KIELIPANKKI
The Language Bank of Finland

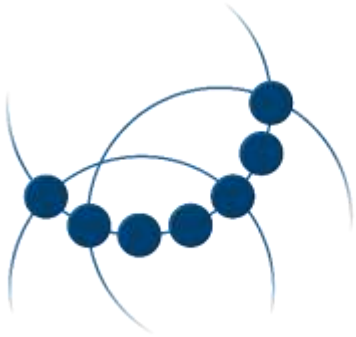
FIN-CLARIAH

Krister Lindén
Research Director of Language Technology
Director of the Language Bank of Finland (Kielipankki)
National Coordinator of FIN-CLARIN
Principal Investigator of FIN-CLARIAH

FIN-CLARIN

UNIVERSITY OF HELSINKI

CLARIN
Common Language Resources and Technology Infrastructure



KIELIPANKKI
The Language Bank of Finland

DARIAH-EU

Digitizing, analysing and visualising data for research in Social Sciences and Humanities

– What is a Research Infrastructure in SSH?

What do SSH vs. ITC need for science?

DARIAH-EU
CLARIN
Common Language Resources and Technology Infrastructure

https://vlo.clarin.eu/?0

Virtual Language Observatory Search Help CLARIN

https://vlo.clarin.eu

CLARIN Virtual Language Observatory

Welcome to the VLO!

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.

See all records | Take a quick tour

about
1.1 million
data sets

Search

Showing all

Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

- Language
- Collection
- Resource type
- Modality
- Format
- Keyword

<< < 1 2 3 4 5 6 7 8 9 10 > >>

EXMARaLDA Demo corpus

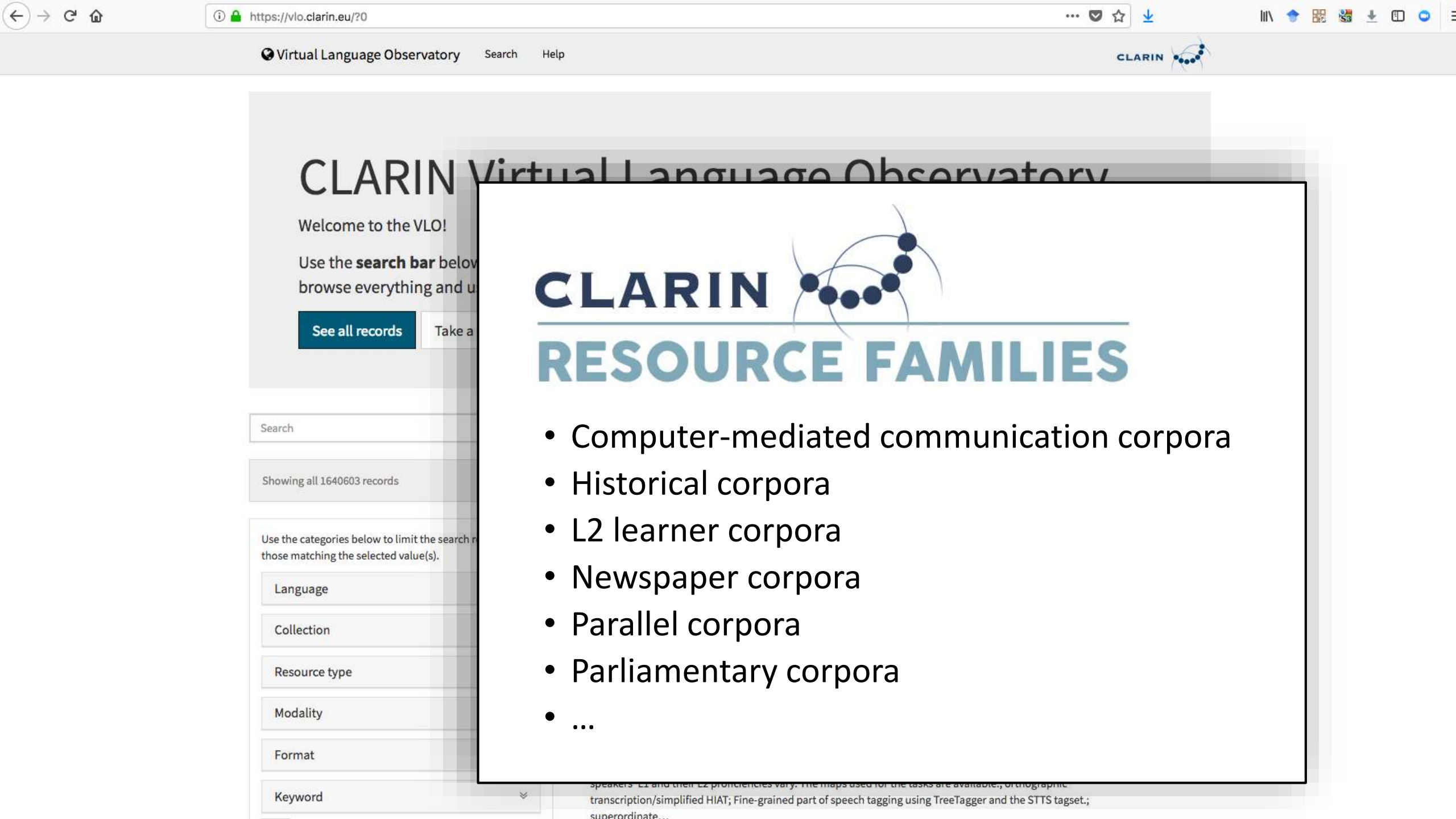
(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

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 information; accentuation/stress; English translation; Standard German translation; German translation; Englisch translation; code-switch

The Hamburg MapTask Corpus (HAMATAC)

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

Audio and two video recordings of map tasks with adult L2 users of German and one L1 speaker. The speakers' L1 and their L2 proficiencies vary. The maps used for the tasks are available.; orthographic transcription/simplified HIAT; Fine-grained part of speech tagging using TreeTagger and the STTS tagset.; superordinate...


https://vlo.clarin.eu/?0
Virtual Language Observatory
Search
Help
CLARIN
CLARIN Virtual Language Observatory
Welcome to the VLO!
Use the search bar below
browse everything and u
See all records
Take a
Search
Showing all 1640603 records
Use the categories below to limit the search r
those matching the selected value(s).
Language
Collection
Resource type
Modality
Format
Keyword
CLARIN
RESOURCE FAMILIES
• Computer-mediated communication corpora
• Historical corpora
• L2 learner corpora
• Newspaper corpora
• Parallel corpora
• Parliamentary corpora
• …

Resource families & WGs

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora , …

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists , …

Tools

- Normalization
- Part-of-speech tagging and lemmatization
- Named entity recognition
- Tools for sentiment analysis , …

FAIR - Findable, Accessible, Interoperable, Reusable

Online Corpora

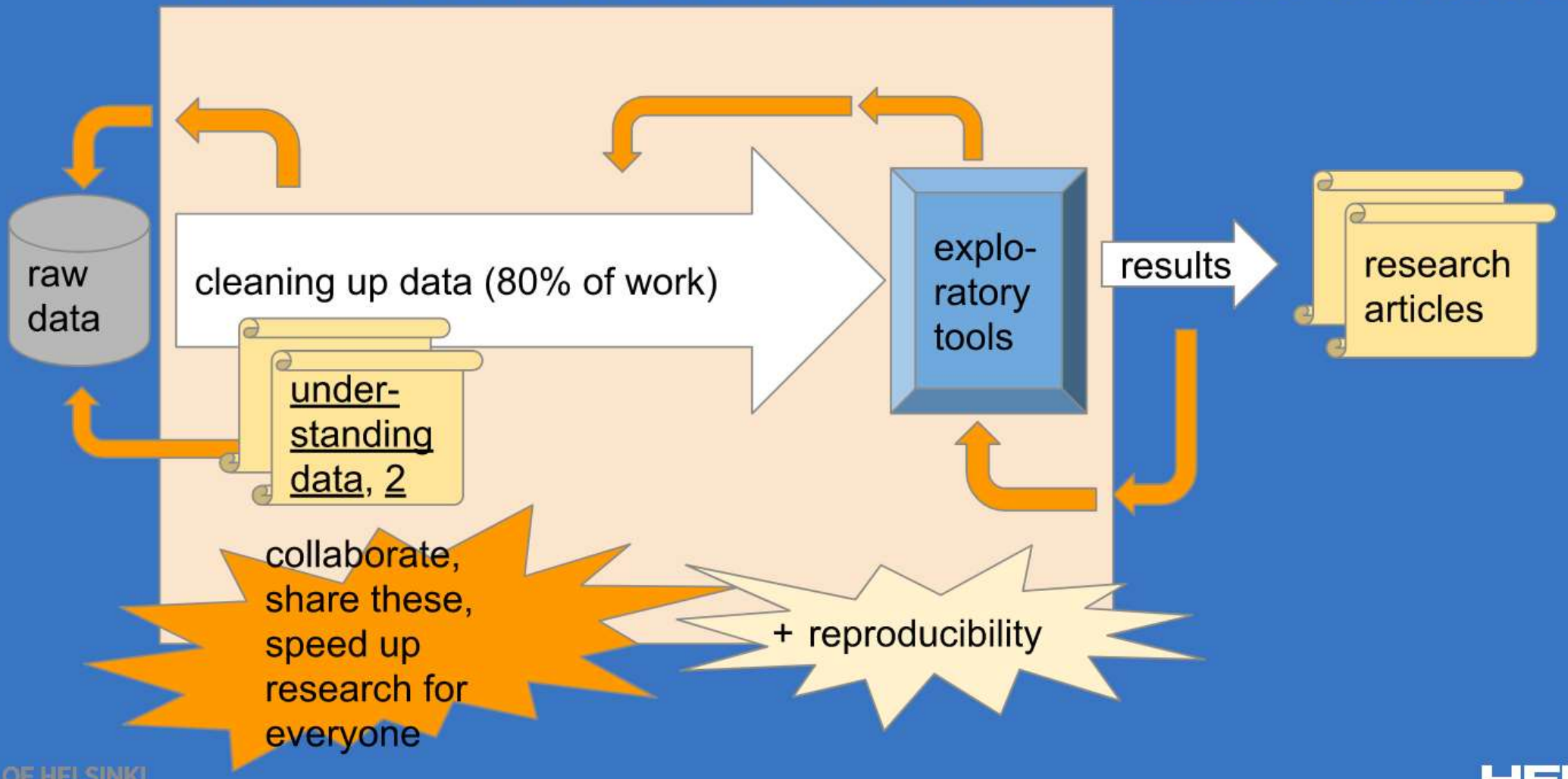
Gw = billion words, Mw = million words, h = hours

Resources	2022
Text	
Magazines and newspapers 1770- (NLF and Web publ.)	15 Gw
Social media and similar sources 2000- (Suomi24, Ylilauta, …)	5 Gw
Literature and manuscripts (Gutenberg, Fennica, archives)	70 Mw
Speech	
Video sessions from the Finnish Parliament 2008-2020	3000 h
Dialect and everyday speech (Kotus, Turku, Donate Speech)	4500 h
Sign language resources (Aalto, Kuurojen liitto)	100 h
Multilingual and Other Resources	
Multilingual Resources (EuroParl, laws, Bible, Opus, …)	5 Gw
Learner's resources (Oulu, Jyväskylä, Kotus, Aalto)	5 Mw
Open source lexicons and terminologies (Helsinki, Tromssa)	400 Kw

Kielipankki has approx. 25 GW in >1400 databases

Resource families in Kielipankki

Leverage collaboration, open science workflows to reduce individual workload



Slide created by Eetu Mäkelä

Cleaning and Exploration

Cleaning tools

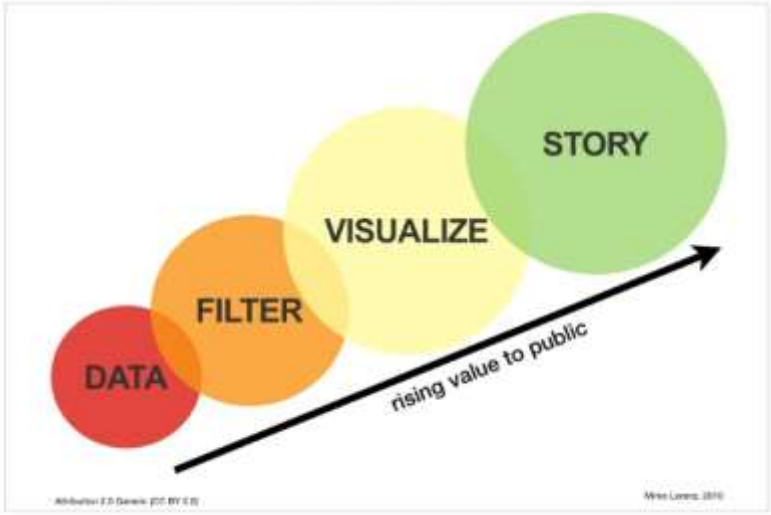
- **Ingestion**
  - Text: Suomi24, web publications, blogs, ...
  - Images: Clay tablets, magazines , manuscripts, …
  - Audio: Radio programs, …
  - Video: Parliamentary and news videos, …
- **Conversion**
  - html2txt
  - OCR & HTR
  - ASR
  - Video2txt
  - …
- **Structuring**
  - Boundary detection: speech act, sentence, paragraph, …

Exploratory tools

- **Annotation**
  - Picture element annotation
  - Language identification
  - Base form annotation
  - Named-entity annotation
  - Linguistic annotation
  - Sentiment & topic annotation
- **Feature aggregation**
  - Feature counting
  - Feature vectors
  - Clustering & linking
- **Overview & visualisation**
  - Concordance
  - Trend diagrams
  - Location maps
  - Topic modelling
  - Network analysis
  - Summarization
- **Reading & understanding**
  - Subset extraction
  - Full-text, picture & video viewers

Infrastructure for digitizing, analysing and visualising data for research in SSH

DARIAH-EU
CLARIN
Common Language Resources and Technology Infrastructure



CLARIN
Common Language Resources and Technology Infrastructure

KIELIPANKKI
The Language Bank of Finland

DARIAH-EU

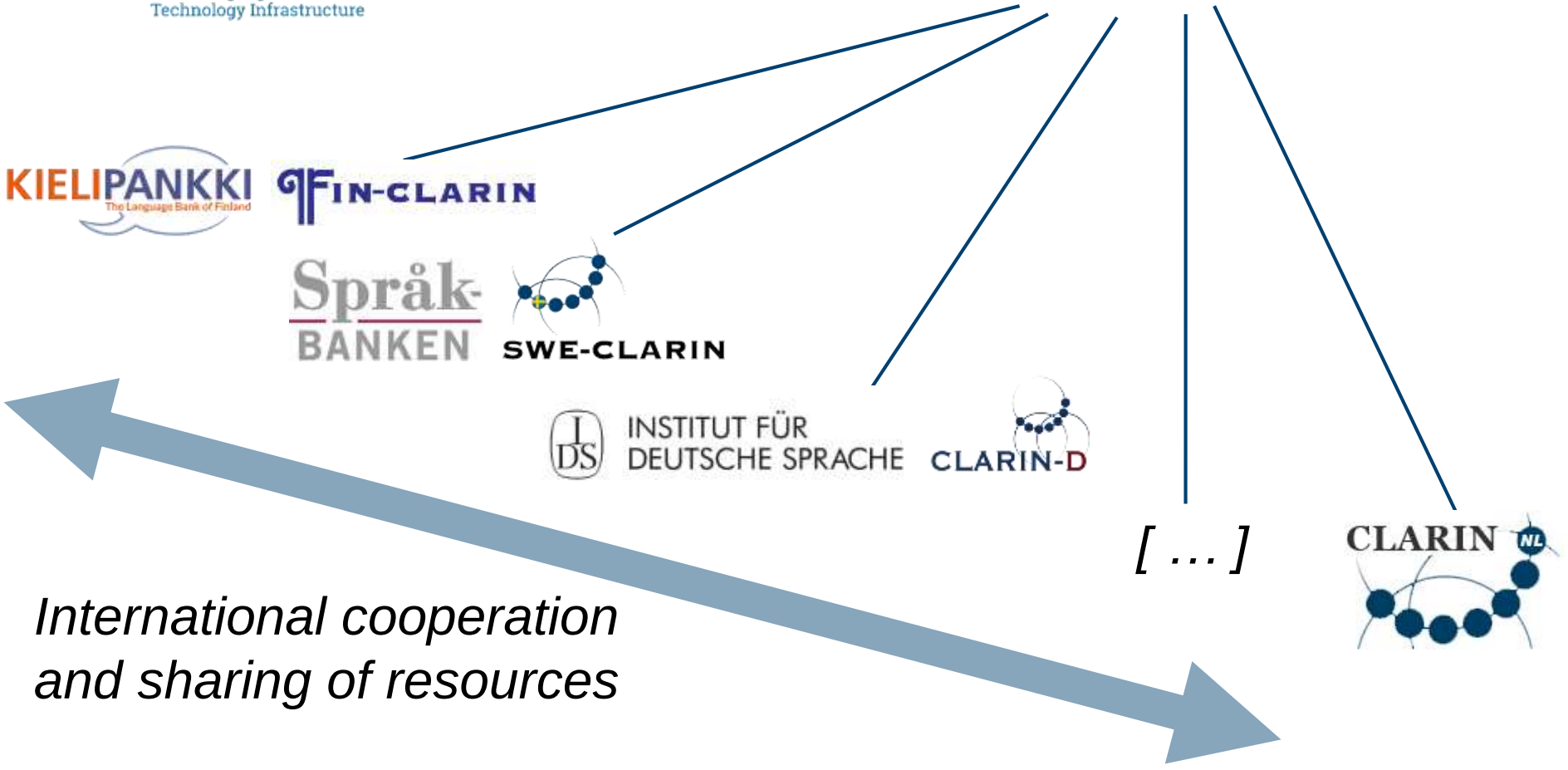
FIN-CLARIAH

International Background

CLARIN ERIC

European Research Infrastructure Consortium
founded on February 29, 2012

https://www.clarin.eu / NL



Member countries:

- **22 members**
- **2 observers**

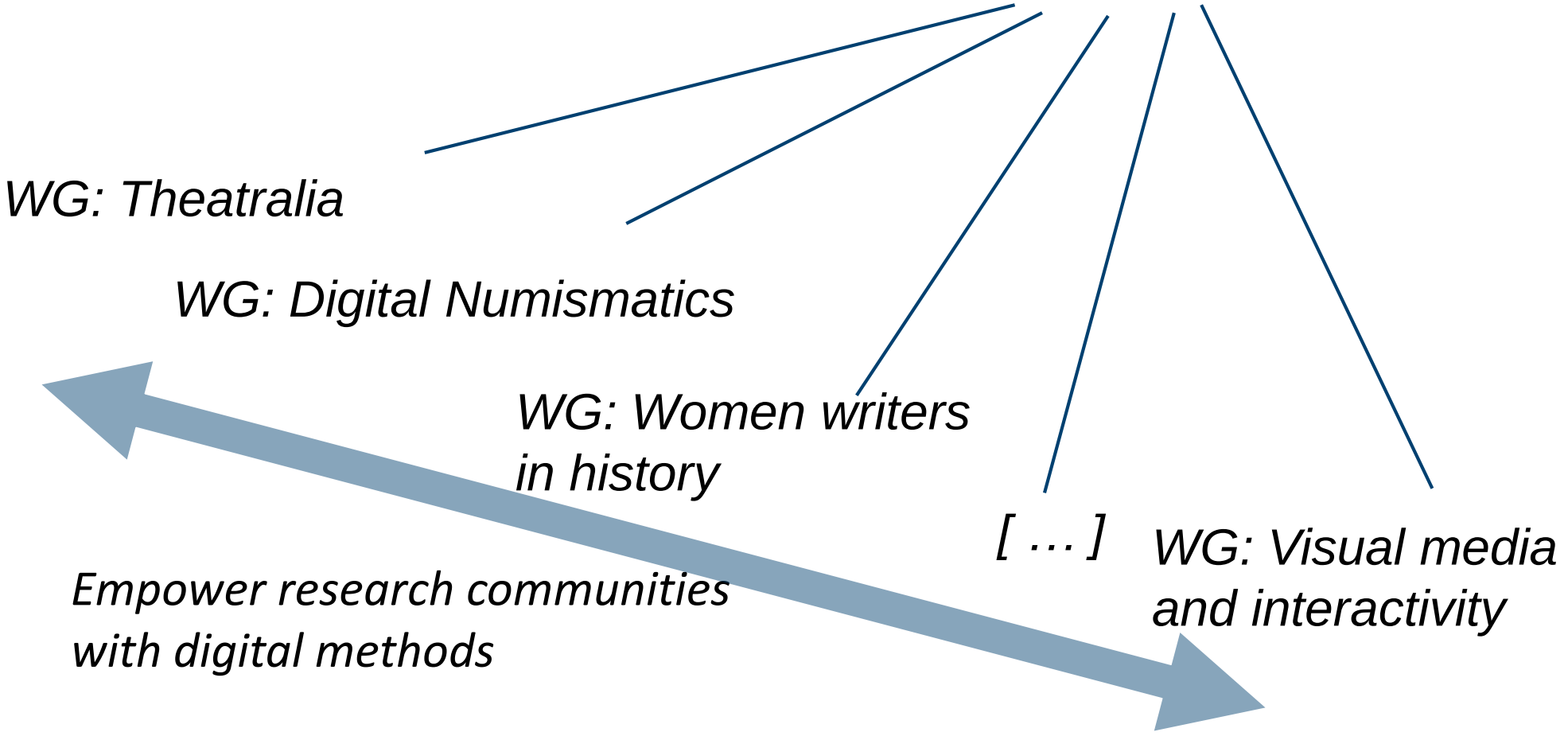
The Netherlands
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Republic
Denmark
Estonia
Finland
Germany
Greece
Hungary
Iceland
Italy
Latvia
Lithuania
Norway
Poland
Portugal
Slovenia
Sweden

Observers:
South Africa
UK
USA / CMU

DARIAH ERIC

European Research Infrastructure Consortium
founded on August 6, 2014

https://www.dariah.eu/



Member countries:
- **19 members**
- **8 coop. partners**

France
Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Republic
Denmark
Germany
Greece
Ireland
Italy
Luxembourg
Malta
Netherlands
Poland
Portugal
Serbia
Slovenia

Coop. partners
Finland, Hungary
Norway, Romania
Slovakia, Sweden
Switzerland
United Kingdom

Resource families & WGs

Corpora

- Computer-mediated communication corpora
- Corpora of academic texts
- Historical corpora
- L2 learner corpora
- Literary corpora
- Manually annotated corpora
- Multimodal corpora
- Newspaper corpora
- Parallel corpora
- Parliamentary corpora
- Reference corpora
- Spoken corpora

Lexical Resources

- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools

- Normalization
- Named entity recognition
- Part-of-speech tagging and lemmatization
- Tools for sentiment analysis

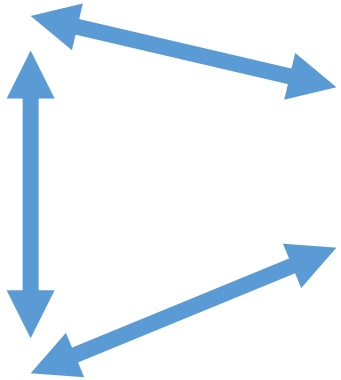
FAIR - Findable, Accessible, Interoperable, Reusable

FIN-CLARIAH – Partners

- University of Helsinki
- CSC – IT Center for Science

Coordination, access to and long-term storage of large **centrally provided resources and tools**

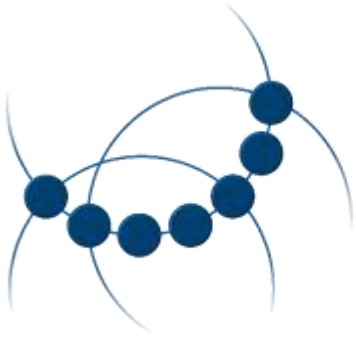
- Aalto University
- Tampere University
- University of Eastern Finland
- University of Jyväskylä
- University of Oulu
- University of Turku
- University of Vaasa
- Institute for the Languages of Finland
- National Archives
- National Library



KIELIPANKKI
The Language Bank of Finland

Provide access to
resources and tools developed locally
by individual researchers or research groups

CLARIN
Common Language Resources and Technology Infrastructure



KIELIPANKKI
The Language Bank of Finland

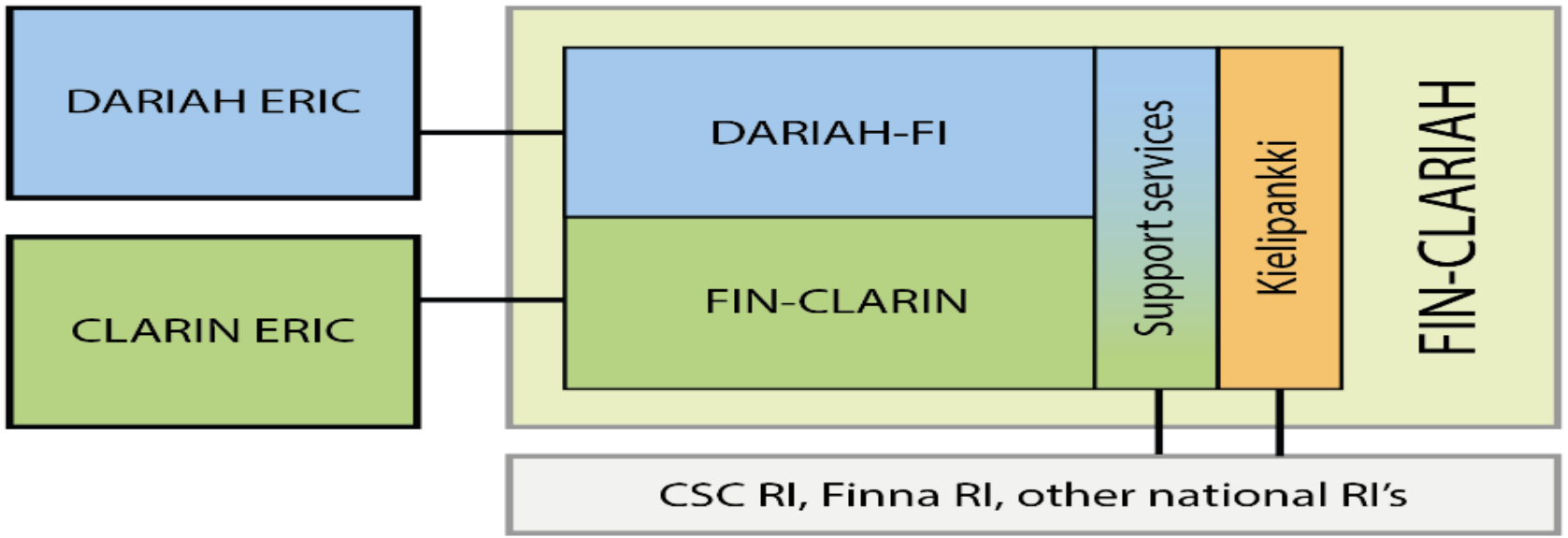
DARIAH-EU

Roadmap 2021-2030

In 2020, FIN-CLARIAH was accepted on the national roadmap comprising the components: FIN-CLARIN and DARIAH-FI

CLARIN
Common Language Resources and Technology Infrastructure

DARIAH-EU



Research Infrastructure for the Humanities and Social Sciences

FIN-CLARIN

Krister Lindén, chair, PI
Mikko Tolonen, vice chair

UNIVERSITY OF HELSINKI

FIN-CLARIAH Roadmap for 2022-2030

DARIAH-EU
CLARIN
Common Language Resources and Technology Infrastructure

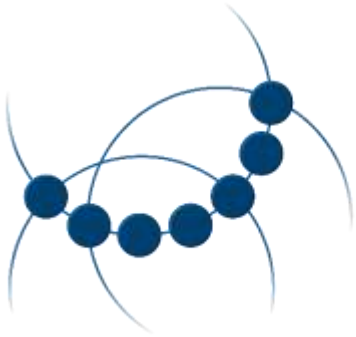
Module 5	Information Interaction (IIA)	UTA
Module 4	Analysing structured data	UHEL
Module 3	Structuring data	UHEL
Module 2	Language Research Infrastructure (LRI)	UHEL
Module 1	Natural Language Processing (NLP)	UHEL

DARIAH-FI

FIN-CLARIN

FIN-CLARIAH

CLARIN
Common Language Resources and Technology Infrastructure



KIELIPANKKI
The Language Bank of Finland

DARIAH-EU

Project 2022-2023

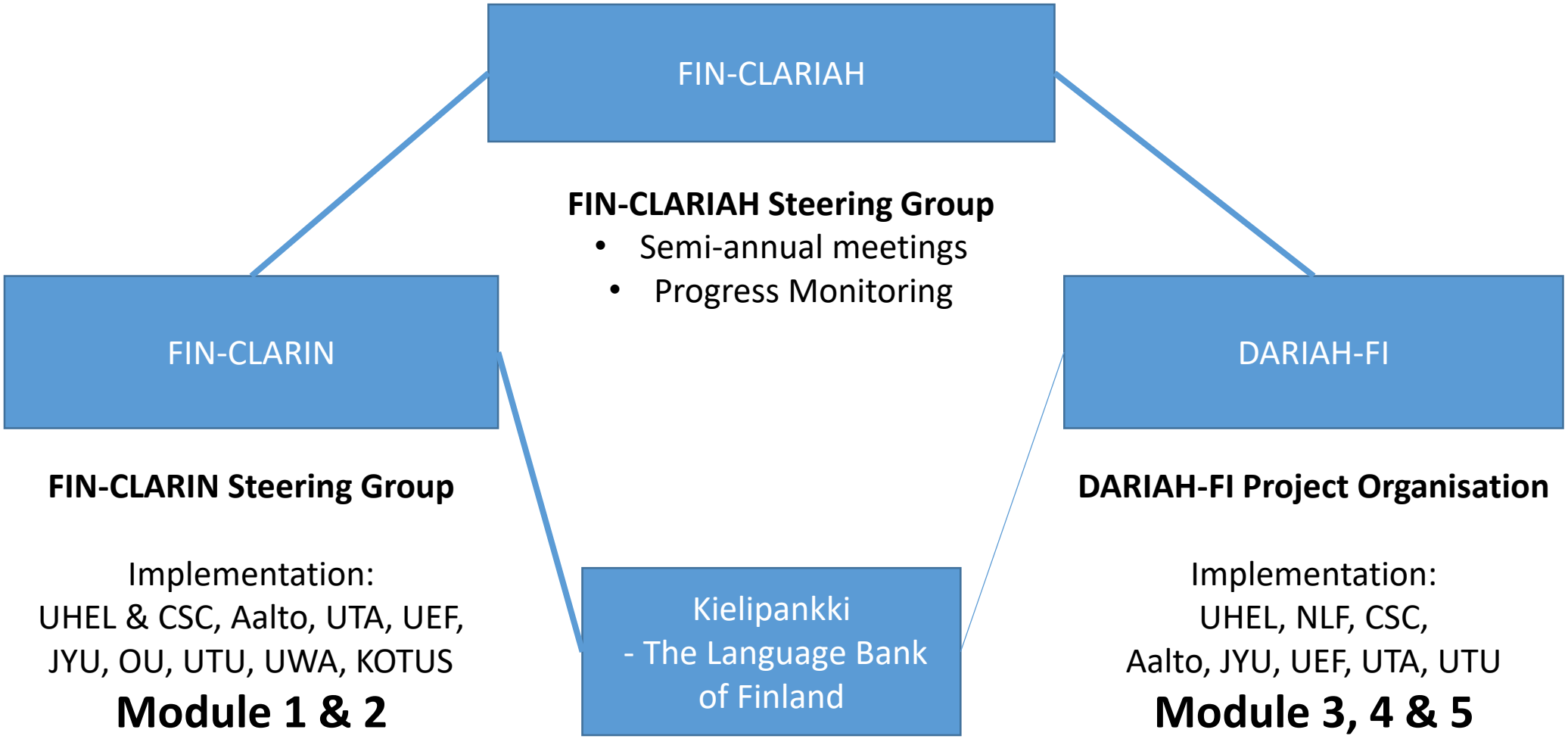
Applied for Academy of Finland infrastructure funding in 2021
for FIN-CLARIAH consolidation

Focus areas

1. Reach beyond processing of **spoken standard Finnish into colloquial speech** (Goal 1) *[FIN-CLARIN]*
2. Cater to a broad range of SSH research needs for processing **unstructured text** (Goal 2) *[FIN-CLARIN/DARIAH-FI]*
3. Facilitate research based on **metadata** (Goal 3) *[DARIAH-FI]*

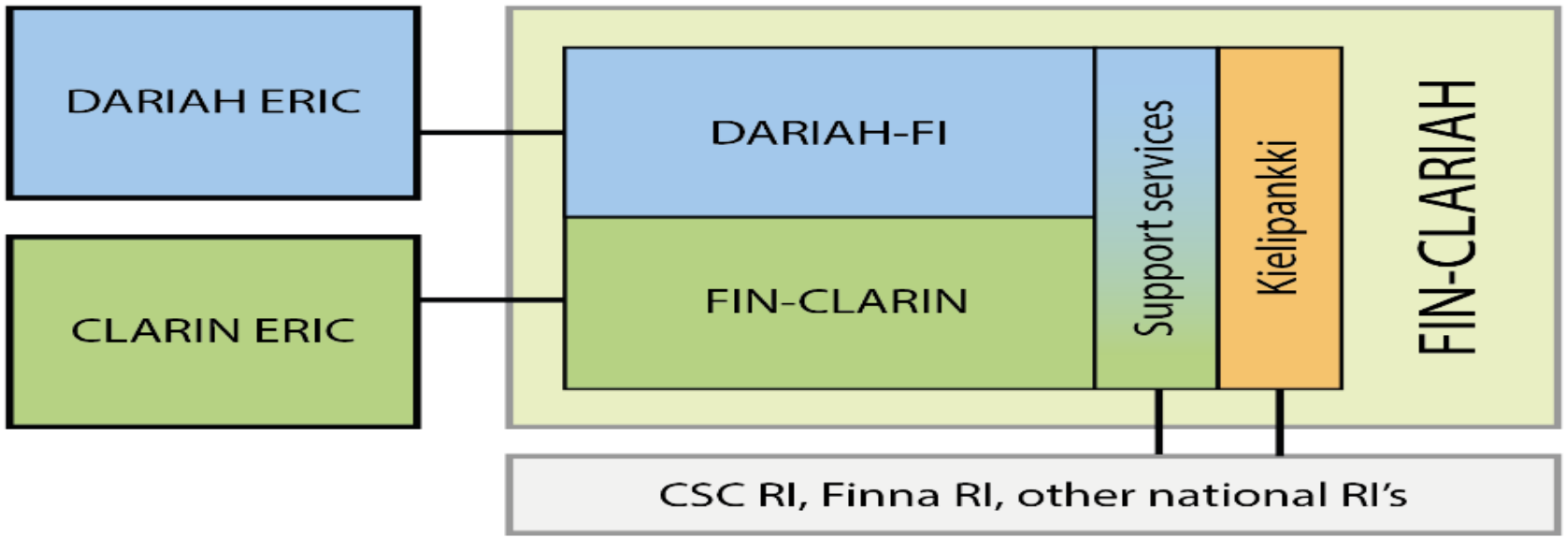
Module 1: Natural Language Processing	Goal 1	Goal 2	Goal 3
Module 2: Language Research Infrastructure	Goal 1	Goal 2	Goal 3
Module 3: Structuring data		Goal 2	Goal 3
Module 4: Analysing structured data		Goal 2	Goal 3
Module 5: Information Interaction	Goal 1	Goal 2	Goal 3

RI organisation



CLARIN
Common Language Resources and Technology Infrastructure

DARIAH-EU



Research Infrastructure for the Humanities and Social Sciences

FIN-CLARIN

Krister Lindén, chair, PI
Mikko Tolonen, vice chair

UNIVERSITY OF HELSINKI