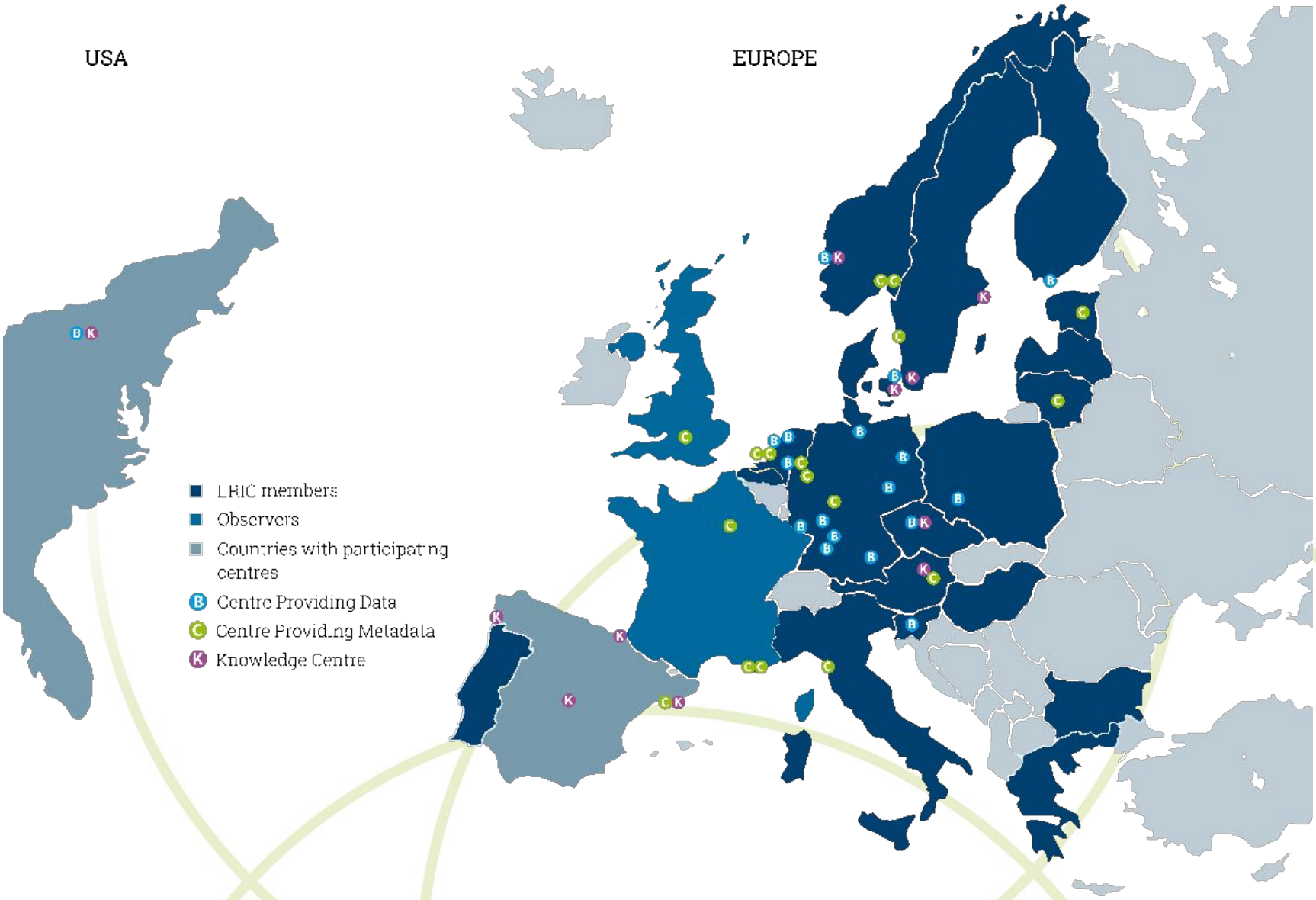
Towards an Open Science Infrastructure for the Digital Humanities: The Case of CLARIN

Koenraad De Smedt , Franciska de Jong , Bente Maegaard,
Darja Fišer and Dieter Van Uytvanck
(CLARIN Board of Directors)

DHN 2018, Helsinki

A European infrastructure

- CLARIN is the European research infrastructure for language resources, primarily aimed at the humanities
- A European Research Infrastructure Consortium (ERIC) with 19 current member countries, 2 observers and 2 additional countries with participating centres
- Active since 2008, ERIC established in 2012, still growing

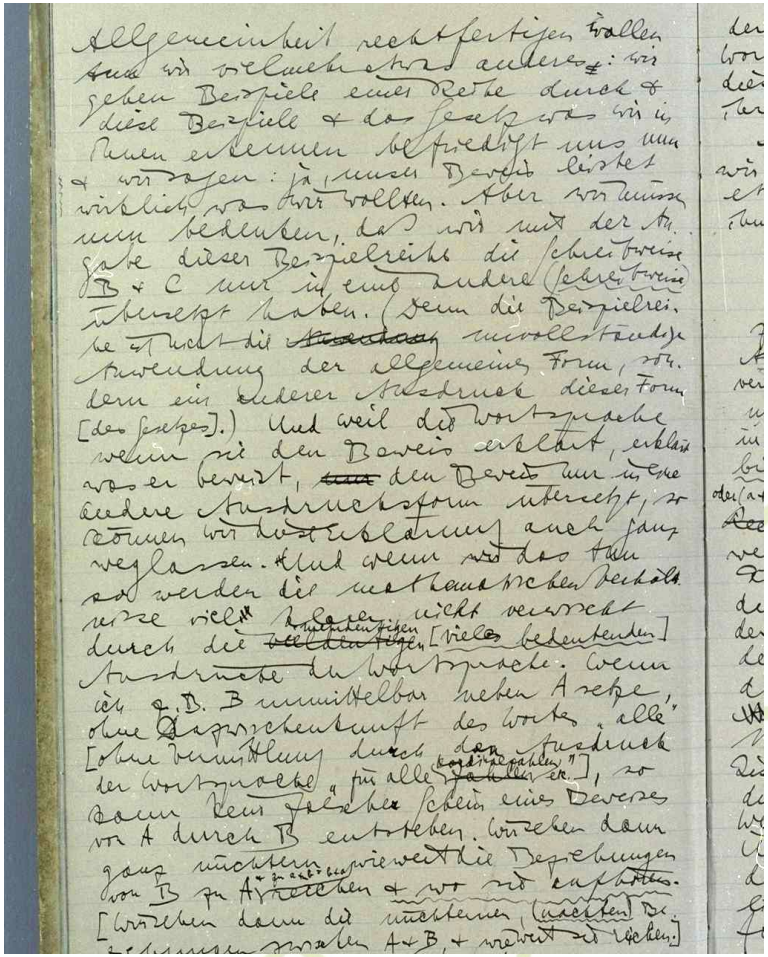

USA
EUROPE
LHIC members
Observers
Countries with participating centres
Centre Providing Data
Centre Providing Metadata
Knowledge Centre

Goal

Not forcing a model on the DH or institutionalizing it, but contributing an infrastructure and meeting ground which aims to make

> “all digital language resources and tools from all over Europe and beyond [...] accessible [...] for the support of researchers in the humanities and social sciences”
>
> (Maegaard et al. 2017)

“Curation, analysis, editing, and modeling comprise fundamental activities at the core of DH.” (Burdick et al. 2012)



Allgemeinheit rechtfertigen **w**ollen
tun wir vielmehr etwas anderes ; <:> wir
gehen Beispiele einer Reihe durch &
diese Beispiele & das Gesetz was wir in
ihnen erkennen befriedigt uns nun
& wir sagen: ja, unser Beweis leistet
wirklich was wir wollten. Aber wir müssen
nun bedenken, daß wir mit der An-
gabe dieser Beispielreihe die Schreibweise
B & C nur in eine andere <(> Schreibweise <)>
übersetzt haben. (Denn die Beispielrei-
he ist nicht die ~~Anwendung~~ unvollständige
Anwendung der allgemeinen Form, son-
dern ein anderer Ausdruck dieser Form
[des Gesetzes] .) Und weil die Wortsprache
wenn sie den Beweis erklärt, erklärt
was er beweist, ~~nur~~ den Beweis nur in eine
andere Ausdrucksform übersetzt, so
können wir diese Erklärung auch ganz
weglassen. **U**nd wenn wir das tun
so werden die mathematischen Verhält-
nisse viel ~~<...>~~ klarer, nicht verwischt
durch die ~~vieldeutigen~~ mehrdeutigen [vie**les** bedeutenden]
Ausdrücke der Wortsprache. Wenn
ich z.B. B unmittelbar neben A
setze, ohne [~~d~~|**D**]azwischenkunft des Wortes „alle”
[ohne Vermittlung durch d[~~as~~|**en**] Ausdruck
der Wortsprache „für alle ~~Zahlen~~ ˇKardinalzahlen < etc. >”] , so
kann kein falsch**e**r Schein eines Beweises
von A durch B entstehen. Wir sehen dann
ganz nüchtern wie weit die Beziehungen
von B zu A ˇ& zu a + b = b + a reichen & wo sie aufhören.
[Wir sehen dann die nüchternen, <(> nackten <)> Be-
ziehungen zwischen A & B, & wie weit sie re<i>chen.]
Man lernt so erst, unbeirrt von

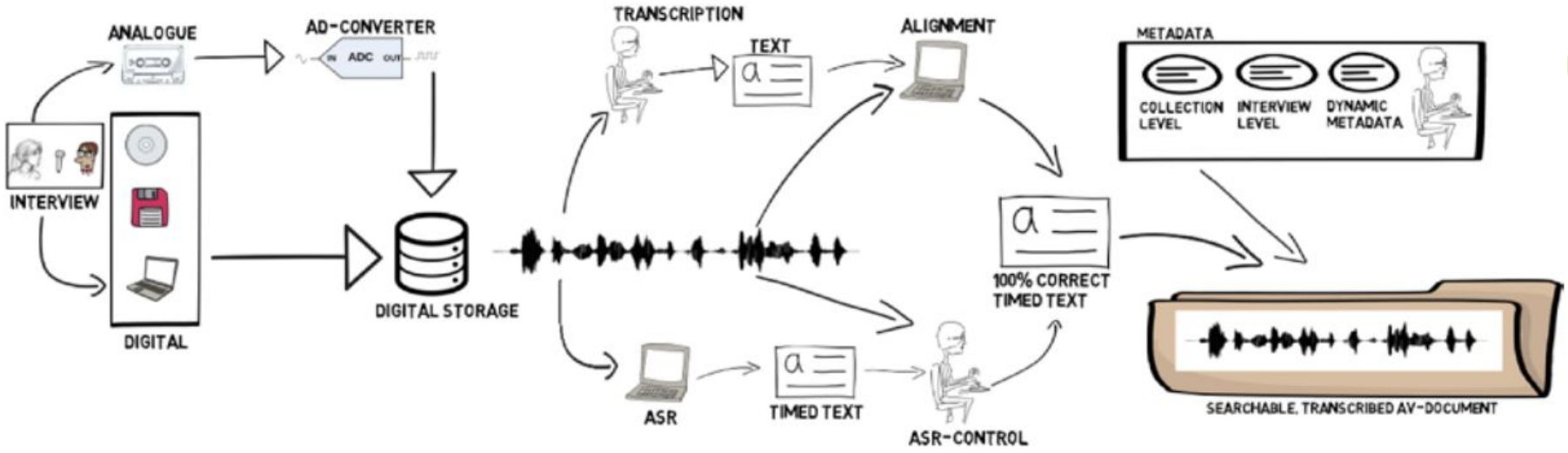
Focus on digital language data curation and access

- Language is the prime source material in the humanities, a rich carrier of human knowledge, emotion, and culture.
- Organized support for digital curation, analysis, editing and modeling involves “platforms, tools, and infrastructures” which “depend upon the basic building blocks of digital activity: digitization, classification, description and metadata, organization, and navigation” (Burdick et al. 2012, *Digital Humanities*)

This is what we do!

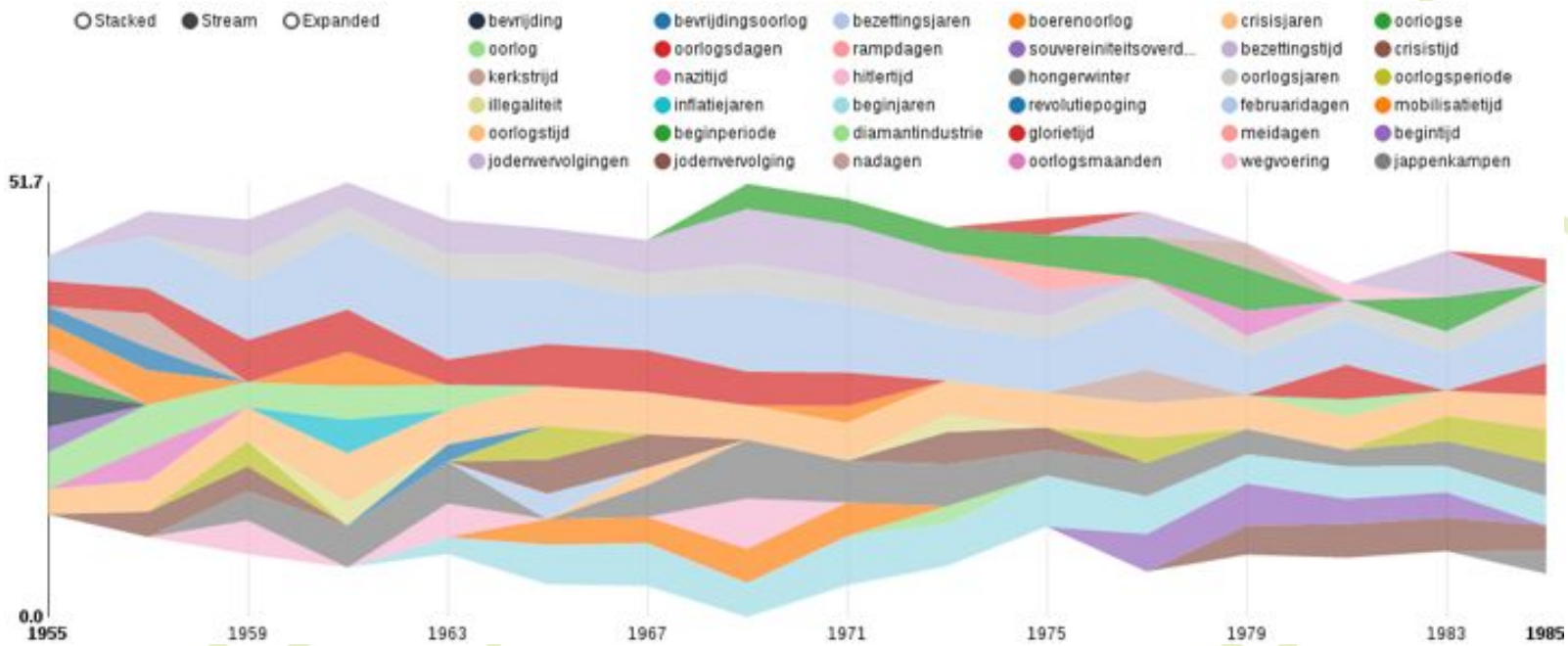
Research questions are the driving forces

- Who was the real author of the Dutch national anthem?
- How was American consumer culture depicted in the Europe throughout the 20th century?
- How can we make a processing chain for curating and preserving oral history?



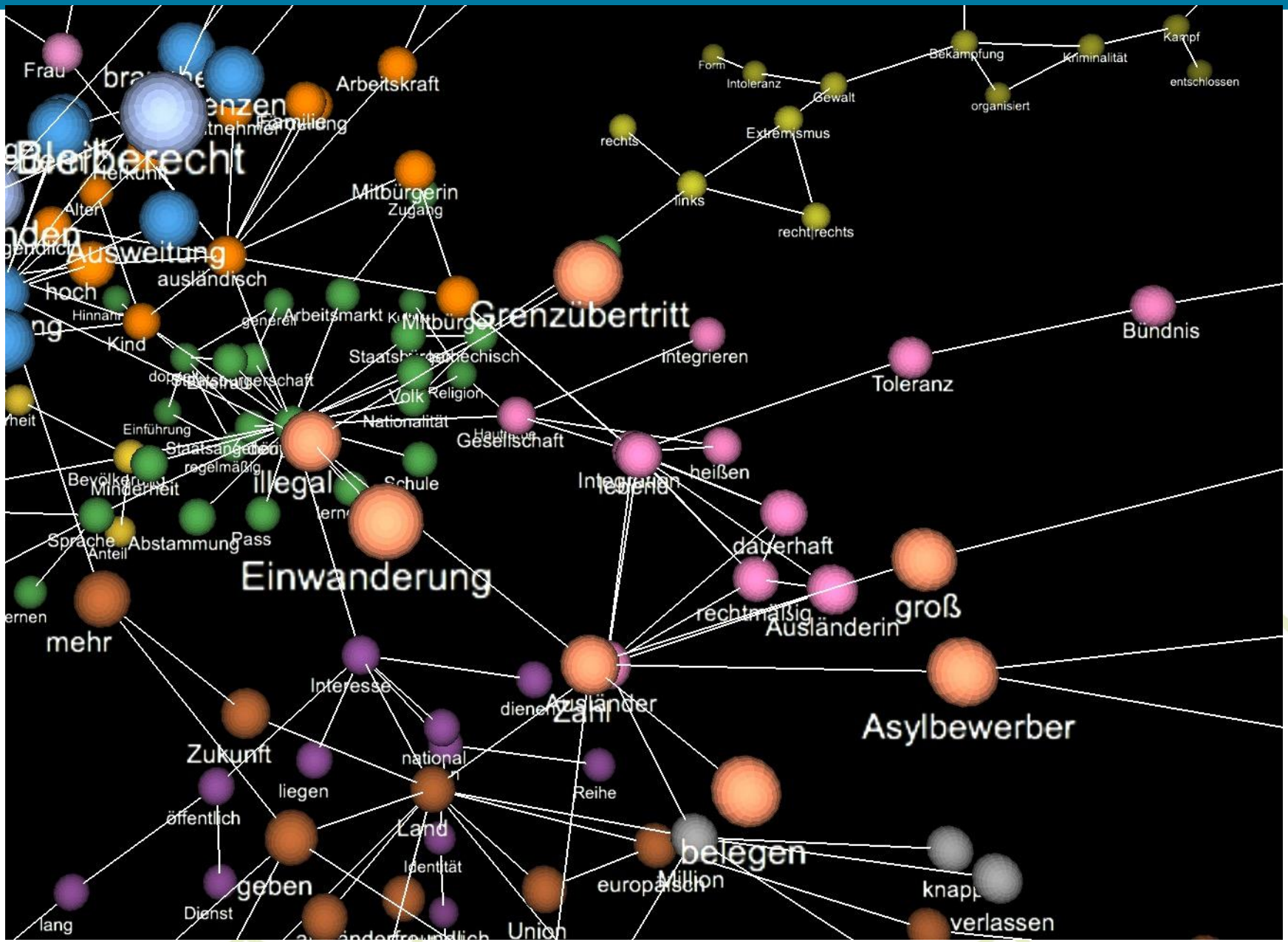
Research questions are the driving forces

- Which changing concepts are associated with *war* in newspapers?



Research questions are the driving forces

- How much polarization is there in social media discourse on climate change?
- Which challenges do language learners face in acquiring grammatical gender in a different language?
- What do historical documents tell us about the relation between gender and work?
- How can we visualize discourse concepts and attitudes by politicians of different parties?



(Andreas Blätte)

CLARIN 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, 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

"CLARIN does not see itself as a stand-alone facility, but rather as 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."

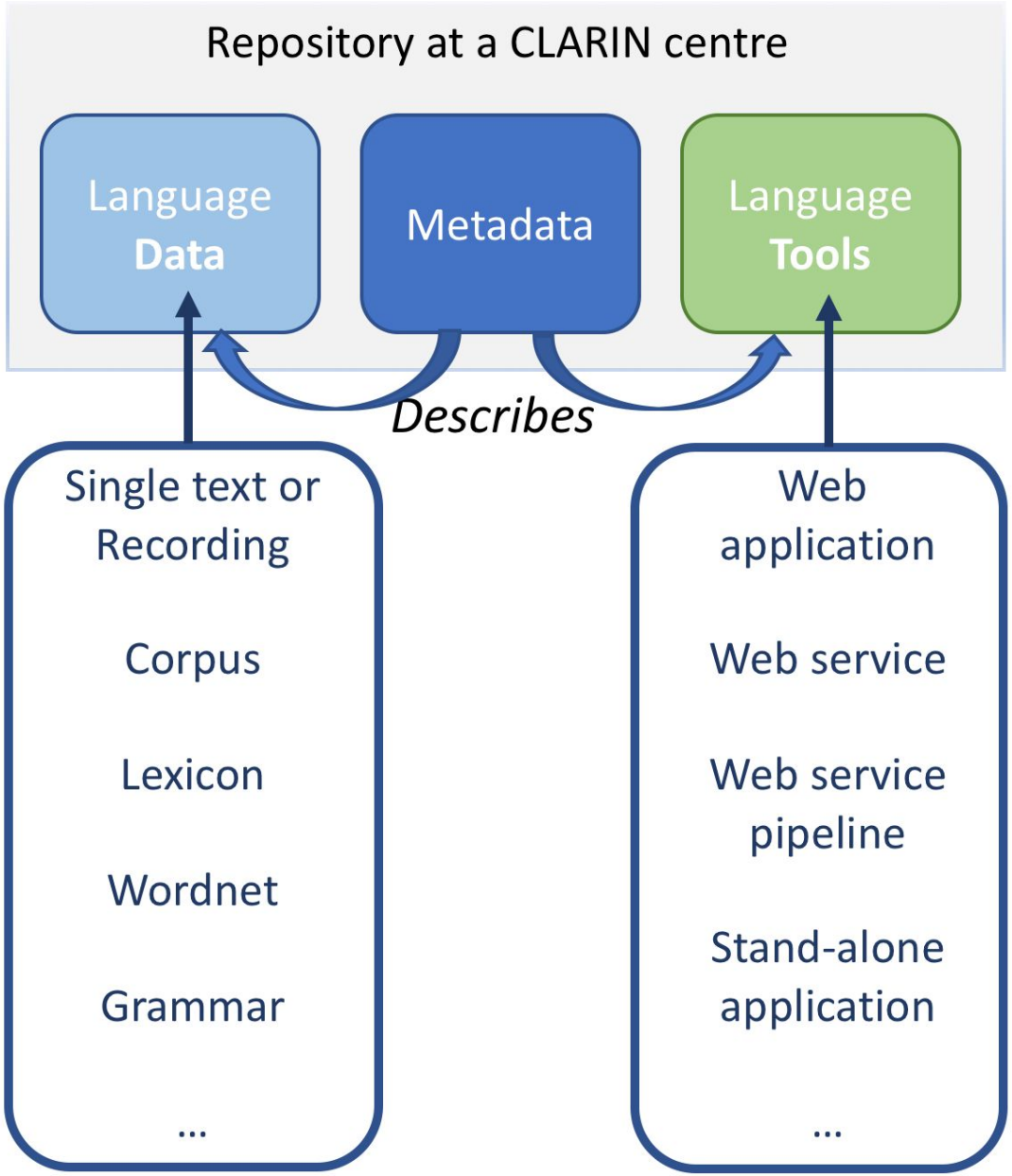
(Maegaard et al. 2017:3)

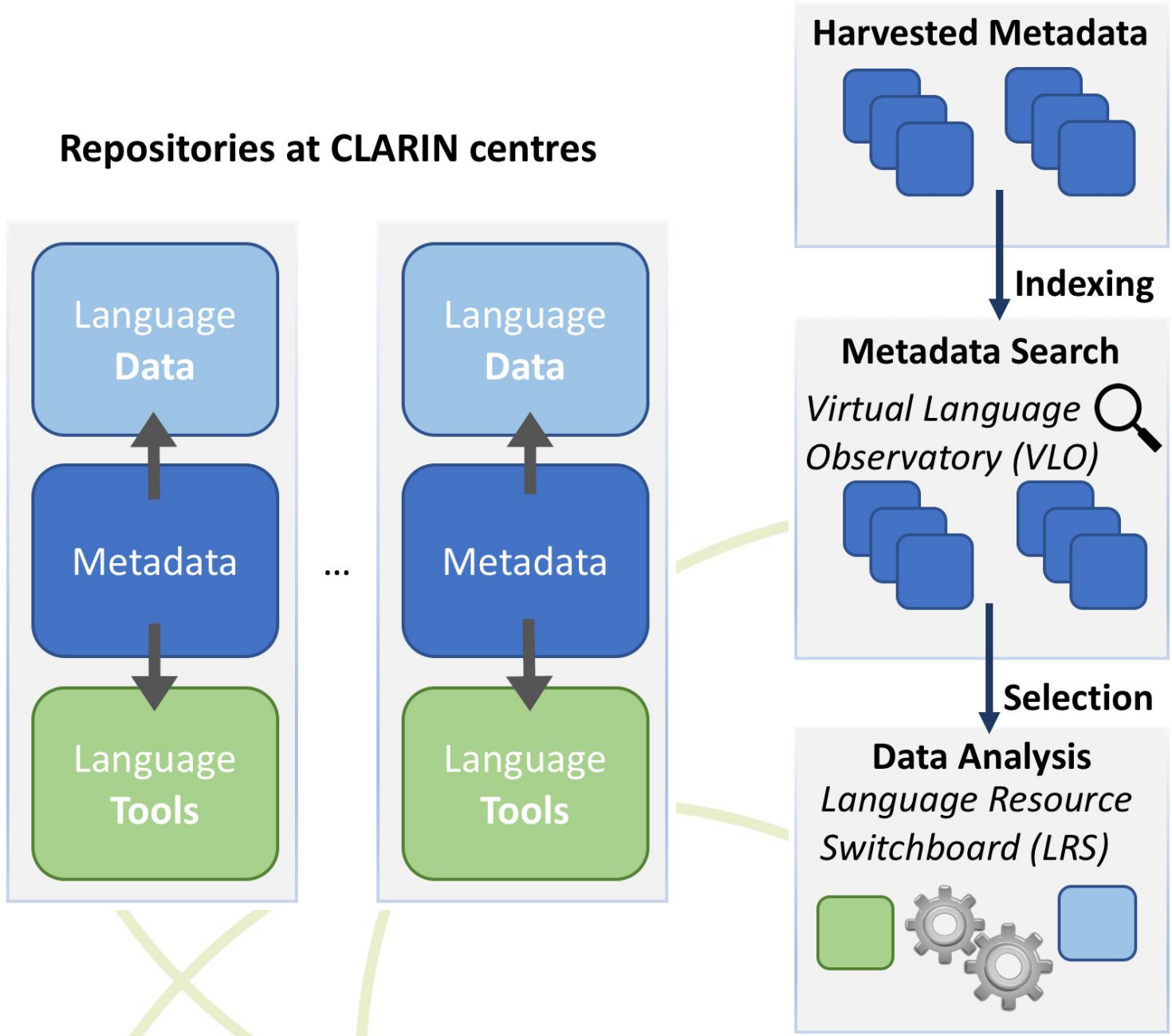
FAIR principles

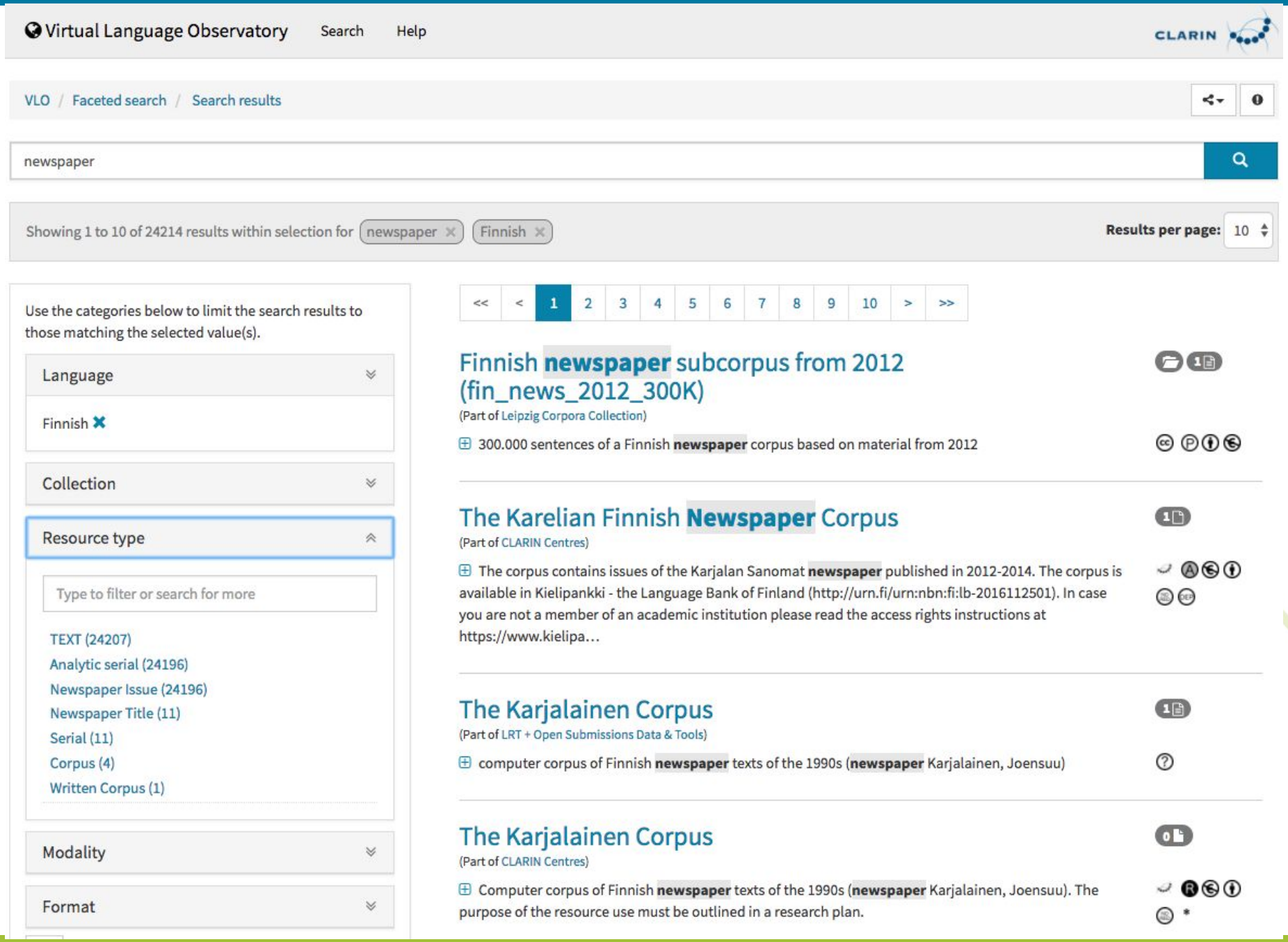
- *Findable:* data must be registered with a persistent ID and items must be collected in a catalog
- *Accessible:* open access protocol (subject to restrictions), clear procedure for authentication and authorization
- *Interoperable:* documented descriptive vocabulary, standards for data and metadata coding
- *Re-usable:* clear licenses, understandable documentation (including provenance), compatibility with community standards and tools

CLARIN catalog

- Virtual Language Observatory (VLO), a registry of Language Resources (LRs) http://vlo.clarin.eu
- 1,600,000 records (including recent addition of Europeana records)
- Component metadata (CMDI, ISO standard)
- Faceted search
- Persistent identifiers for data objects


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)


Virtual Language Observatory
Search
Help
CLARIN
VLO / Faceted search / Search results
newspaper
Showing 1 to 10 of 24214 results within selection for newspaper Finnish
Results per page: 10
Use the categories below to limit the search results to those matching the selected value(s).
Language
Finnish
Collection
Resource type
Type to filter or search for more
TEXT (24207)
Analytic serial (24196)
Newspaper Issue (24196)
Newspaper Title (11)
Serial (11)
Corpus (4)
Written Corpus (1)
Modality
Format
<< < 1 2 3 4 5 6 7 8 9 10 > >>
Finnish newspaper subcorpus from 2012 (fin_news_2012_300K)
(Part of Leipzig Corpora Collection)
300.000 sentences of a Finnish newspaper corpus based on material from 2012
The Karelian Finnish Newspaper Corpus
(Part of CLARIN Centres)
The corpus contains issues of the Karjalan Sanomat newspaper published in 2012-2014. The corpus is available in Kielipankki - the Language Bank of Finland (http://urn.fi/urn:nbn:fi:lb-2016112501). In case you are not a member of an academic institution please read the access rights instructions at https://www.kielipa…
The Karjalainen Corpus
(Part of LRT + Open Submissions Data & Tools)
computer corpus of Finnish newspaper texts of the 1990s (newspaper Karjalainen, Joensuu)
The Karjalainen Corpus
(Part of CLARIN Centres)
Computer corpus of Finnish newspaper texts of the 1990s (newspaper Karjalainen, Joensuu). The purpose of the resource use must be outlined in a research plan.

But how will researchers use it?

"... are the barriers to entry that 'outsiders' perceive real usability issues, or simply points on DH's learning curve?"
(Edwards 2012, p. 213)

1. Improvement of usability of data and tools: online services with good interfaces and visualizations
2. Increase of user capabilities
   a. Knowledge Sharing Infrastructure: K-centres, mobility actions
   b. User involvement actions: surveys, improvement of visibility, master classes, summer school courses, other training events, "Tour de CLARIN"
   c. Website with information, showcases
   d. Workshops that bring together researchers in focused communities

Responsible Data Science

Fair, Accurate, Confidential and Transparent (FACT)

Big data are relevant for DH (e.g. distant reading, social media mining, network analysis etc.)

- Provenance included in metadata
- Clear licensing practice (with CLARIN license categories)
- Analysis tools recommended by CLARIN Language Resource Switchboard

Conclusion and final remarks

- CLARIN is an infrastructure promoting and supporting Open Science for the Digital Humanities
- Wide scope: languages, disciplines, countries, cultures, historical time spans
- Participation in EU projects (such as EOSC-hub) and cooperation with other infrastructures
- Starting cross-sector collaboration with the GLAM sector (Galleries, Libraries, Archives, Museums)

This infrastructure is here for you to use!

(For full references see the paper)


CLARIN
http://clarin.eu