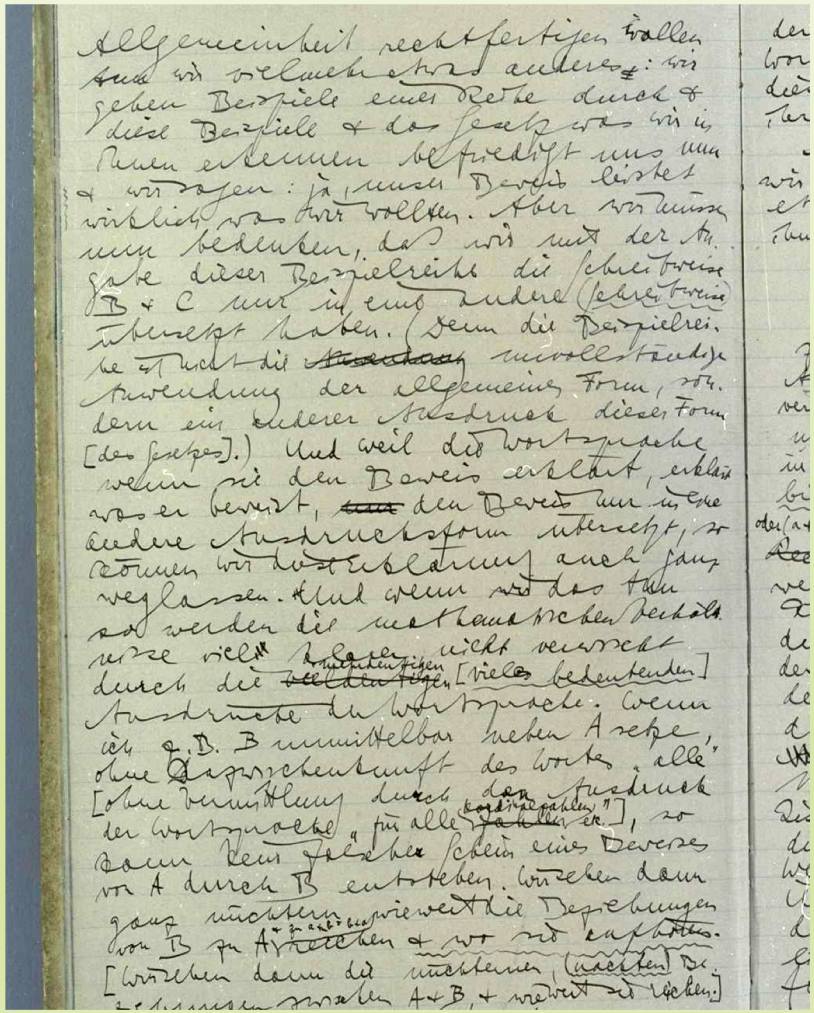
FAIR data for research: the case of language data and tools

Koenraad De Smedt
LLE / CLARINO

UiB, Seminar on Open Access and Open Data
Open Access Week 2018, Oct. 24

Curation, analysis, editing, and modeling comprise fundamental activities at the core of the Digital Humanities.
(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 [viel**es** bedeutenden]
Ausdrücke der Wortsprache. Wenn
ich z.B. B unmittelbar neben A
setze, ohne [~~d~~|**D**]azwischenkunft des Wortes „alle"
[ohne Vermittlung dur**c**h 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

Corpuscle :: Aviskorpus ann. :: Concordance

Basic search | **Advanced search**

"(i|på)" "Austevoll" %c

Run Query | Saved queries | Save

Done. Real time: 0.3853 sec. (0.02

Hit 1 - 22 of 22 | Download (Excel mode) | Type: kwic | Show line filter | Attributes ... | Structures ... | Page si

	match	
1,85 i promille etter ulykken. Fakta: Claus Lundekvam (36) Født: 22. februar 1973	i Austevoll	Tidligere klubber: Selbjørn IL, Brann
1886 har vi aldri hatt besøk av kongelige, sier ordfører Helge Andre Njåstad (Frp)	i Austevoll	. Det er likevel ikke første gang en k
em og blir servert et omfattende kulturprogram. På kvelden tirsdag blir det middag	i Austevoll	. Dit er også ordførere fra alle nabok
. Se hele videointervjuet her Fakta: Claus Lundekvam (36) Født: 22. februar 1973	i Austevoll	Tidligere klubber: Selbjørn IL, Brann
Fiskeformuende Brødrene Ole Rasmus og Helge Arvid Møgster har lignende effekt	på Austevoll	kommune i Hordaland. De to fiskerie
odelljobber. Nå vil han vinne «Skal vi danse». Ta en Travolta. Den gangen hjemme	i Austevoll	dro han det litt for langt. - Jeg reiste
med mann og tre born i fjor. Dei hadde lyst til å prøve eit anna liv, leigde ut huset	i Austevoll	og reiste til Vadsø. Mannen hadde få
nger sør finst det ein annan ordførar frå Framstegspartiet. Helge André Nistad (24)	i Austevoll	er den yngste ordføraren i landet. De
alla ein bananrepublikk. Her er fiskeriet alfa og omega, og kvar tjuande innbyggjar	i Austevoll	er millionær. Næringslivet er variert,
til Askøy: Der hadde me ein Kvalvåg i nordaust i Lindås og to slike vågar mot sør –	i Austevoll	og Sveio. Dessutan Kvalvågen i Aust
ellet. Møgster-familiane – Har de ikkje vore klare over at det er to Møgster-familiar	på Austevoll	? – Jau, vi har ikkje bomma på det.
Vibeke er 27 år, har eigen gåvebutikk og er leiar for kjøpesenteret her i Bekkjarvik	på Austevoll	. Ho er modelltynn som det høver se
emiske danningskulturen er eit trugsmål mot dei livsviktige demokratiske verdiane	på Austevoll	. Kunsten å fortøya ein fiskebåt er vi
-å Pareto i Oslo er her i forhandlingar med DOF, det største og viktigaste selskapet	på Austevoll	. Leiaren har sjølvsagt same etterna
Lakseavfall ble millionbedrift Lakseslo gjev pengar i kassa hjå Hordafor	på Austevoll	i Hordaland. Bedrifta som først byrja
å 95 millionar. Selskapet har rundt 50 tilsette. Dei fleste er knytt til hovudkontoret	på Austevoll	, men selskapet har òg avdelingar på
J. Sjølv er eg oppvaksen på kysten, med klart fotfeste i ei fiskarslekt med utspring	i Austevoll	. Eg bur i eit område der oppdrettsn
on til vindkraftutbygging i stor målestokk i Sunnhordland, sjølv om kommunestyret	i Austevoll	sa samrøysta nei. Etter mi meining e
ett mot vindmøller. Partiet viste til konsesjonen for bygginga av Selbjørn kraftverk	i Austevoll	kommune som vart gjeve trass i at e
Stortinget måndag. Partiet viste til konsesjonen for bygginga av Selbjørn kraftverk	i Austevoll	kommune som vart gjeve trass i at e
U ein lokal variant av same innlegg, der dei peikar ut Midtfjellet i Fitjar og Selbjørn	i Austevoll	som «gode» prosjekt. NU sin hovudp
: underleg lys når vi ser på kva partiet gjer lokalt. Lokalt har Frp støtta bompengar	i Austevoll	og Os der partiet endåtil har ordføra

Even if data is declared open, it cannot always be used the way researchers would like.

What is the use of data if it ...

1. Cannot be found?
   ERROR 404: Page not found
2. Is not accessible?
   Access denied
3. Is incompatible with other data and tools?
   Unrecognized format
4. Cannot be reused?
   Documentation / License not available

Some goals beyond open data

- Make it possible to assess the quality of research
- Make research replicable
- Allow reuse of data in innovative research and development

FAIR: Requirements for data-driven research

Data should be ...

Findable

Accessible

Interoperable

Reusable

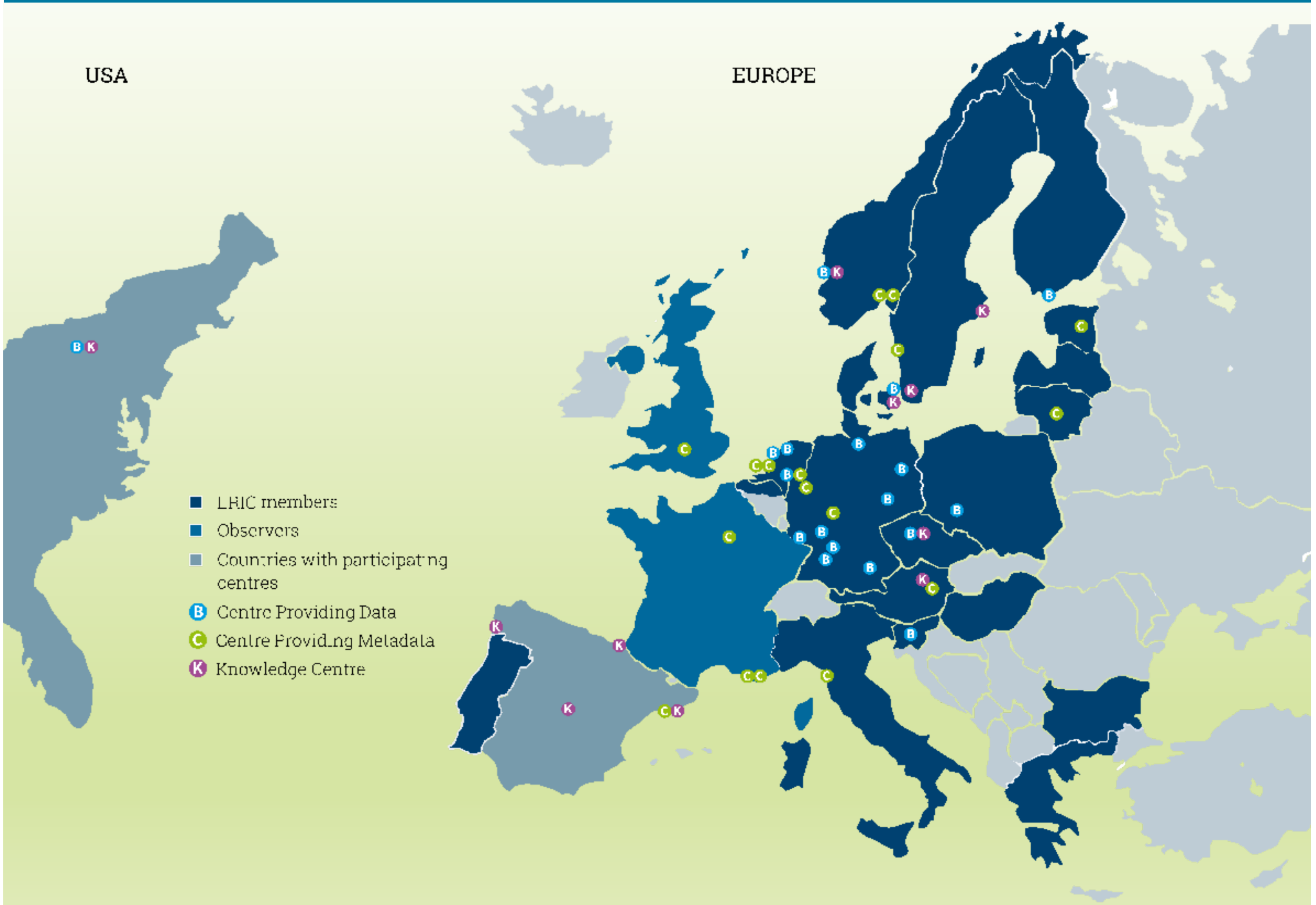
(FAIR Data Principles, Wilkinson et al. 2016)

What do the FAIR principles entail?

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

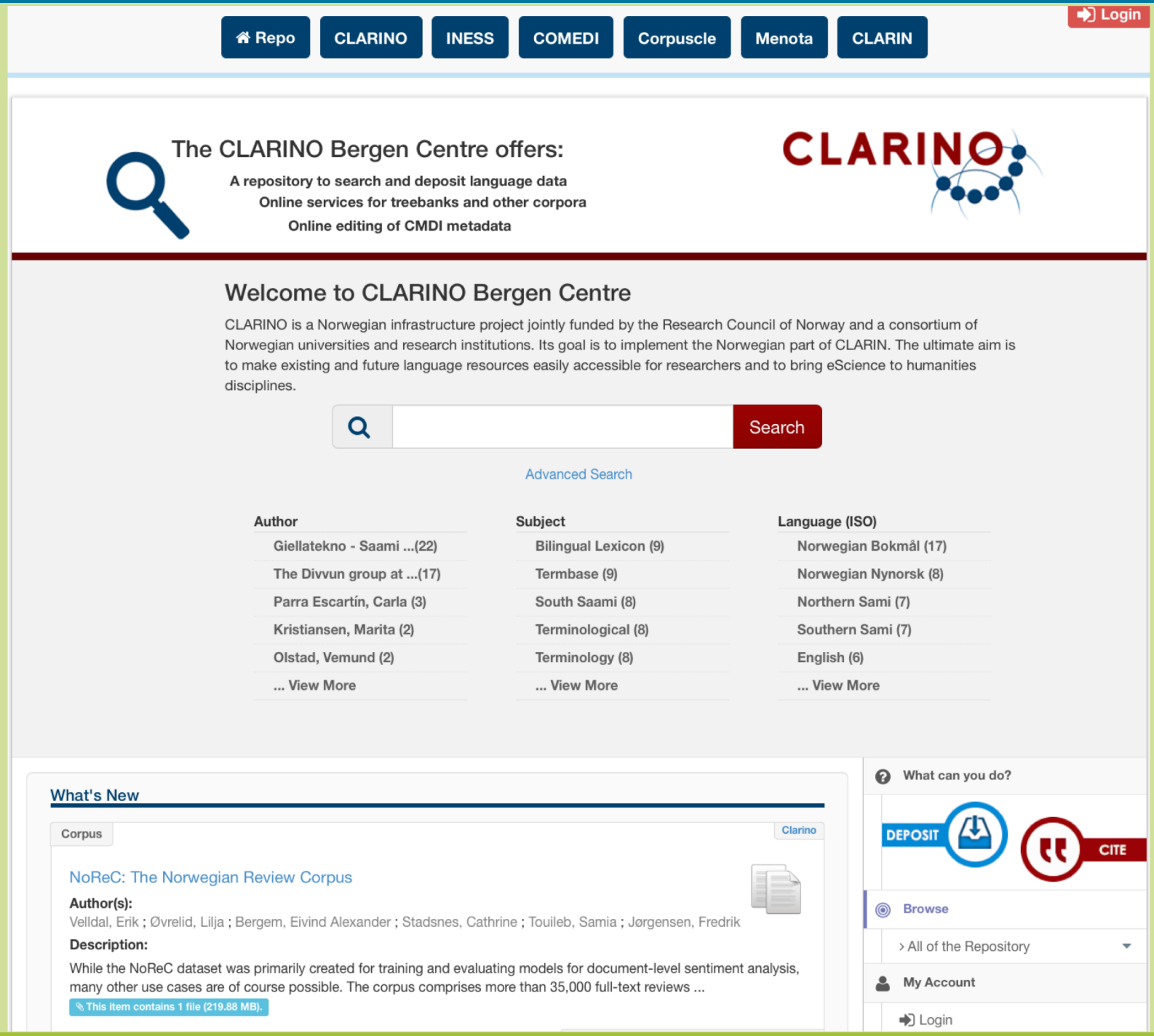
CLARIN: a European infrastructure

- CLARIN is the European research infrastructure for language resources and technology, primarily aimed at the humanities
- A European Research Infrastructure Consortium (ERIC) with 20 current member countries, 2 observers and 2 additional countries with participating centres
- Norway joined in 2015


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

CLARIN in Norway

1. Centres with repositories
    1. UiB/UB (45 downloadable datasets)
    2. UiO/Tekstlab (16 downloadable datasets and 2 tools)
    3. NB/Språkbanken (42 downloadable datasets)
    4. UiT/Trolling (60 downloadable datasets)
    5. UiT/Giellatekno
2. National catalog at the National Library of Norway which harvests metadata from other centres


Repo
CLARINO
INESS
COMEDI
Corpuscle
Menota
CLARIN
Login
The CLARINO Bergen Centre offers:
A repository to search and deposit language data
Online services for treebanks and other corpora
Online editing of CMDI metadata
CLARINO
Welcome to CLARINO Bergen Centre
CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines.
Search
Advanced Search
Author
Giellatekno - Saami ...(22)
The Divvun group at ...(17)
Parra Escartín, Carla (3)
Kristiansen, Marita (2)
Olstad, Vemund (2)
... View More
Subject
Bilingual Lexicon (9)
Termbase (9)
South Saami (8)
Terminological (8)
Terminology (8)
... View More
Language (ISO)
Norwegian Bokmål (17)
Norwegian Nynorsk (8)
Northern Sami (7)
Southern Sami (7)
English (6)
... View More
What's New
Corpus
Clarino
NoReC: The Norwegian Review Corpus
Author(s):
Velldal, Erik ; Øvrelid, Lilja ; Bergem, Eivind Alexander ; Stadsnes, Cathrine ; Touileb, Samia ; Jørgensen, Fredrik
Description:
While the NoReC dataset was primarily created for training and evaluating models for document-level sentiment analysis, many other use cases are of course possible. The corpus comprises more than 35,000 full-text reviews ...
This item contains 1 file (219.88 MB).
What can you do?
DEPOSIT
CITE
Browse
› All of the Repository
My Account
Login

CLARINO

UiO : Institutt for lingvistiske og nordiske studier
Det humanistiske fakultet

tekstlab

Home CLARINO CLARIN Text Laboratory

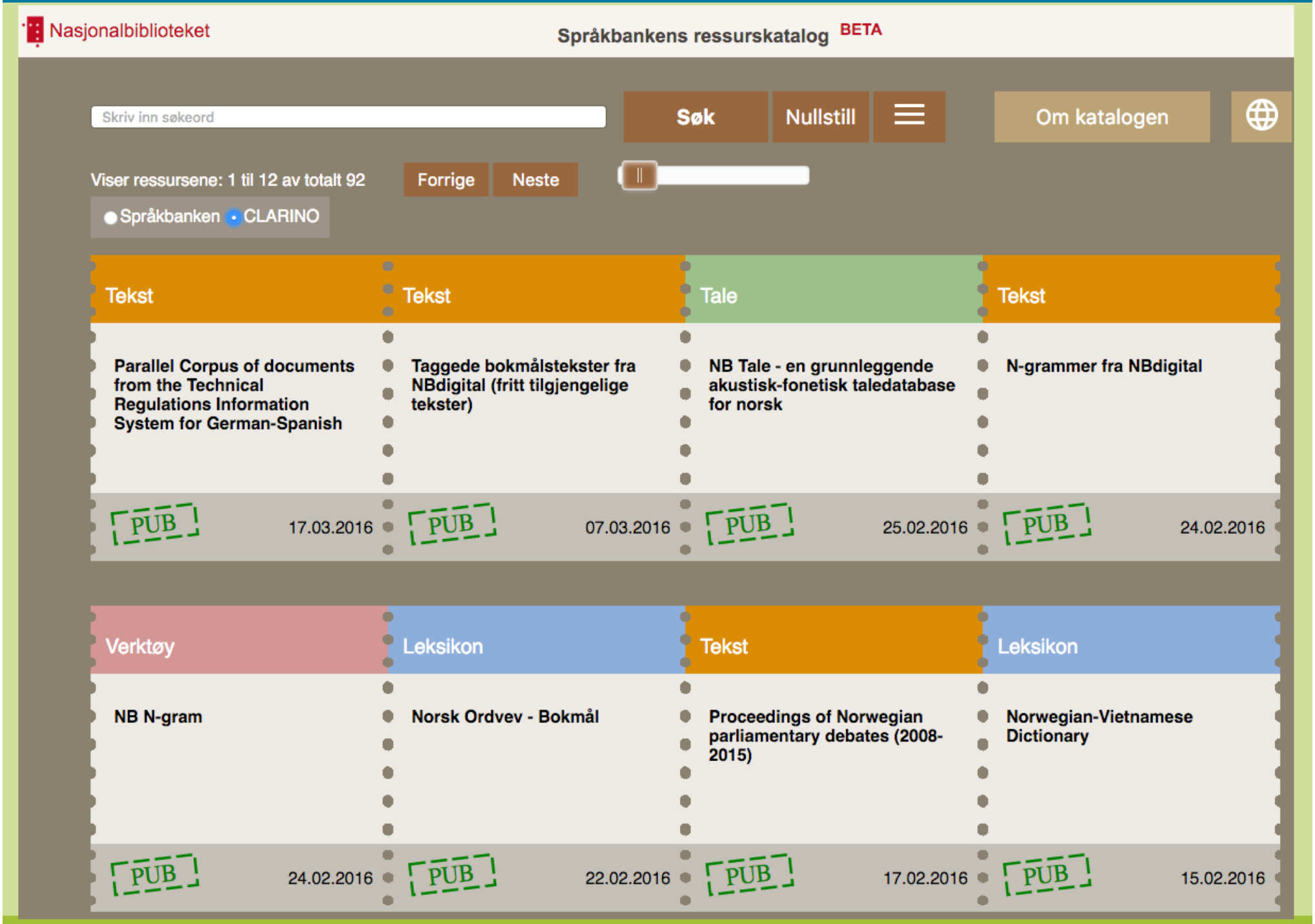
Welcome to CLARINO Text Laboratory Centre

CLARINO is a Norwegian infrastructure project jointly funded by the Research Council of Norway and a consortium of Norwegian universities and research institutions. Its goal is to implement the Norwegian part of CLARIN. The ultimate aim is to make existing and future language resources easily accessible for researchers and to bring eScience to humanities disciplines. The CLARINO project is coordinated by University of Bergen.

CLARINO Text Laboratory Centre is a C centre in the CLARIN infrastructure.
The table below shows Text Laboratory resources with a signed CLARIN agreement. More resources will come. Go to the Text Laboratory homepage to view all resources from the Text Laboratory.

Corpora:

The Big Brother Corpus	(2007) 550 000 words. Speech. Norwegian TV show from 2001. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Get username and password - Search the corpus
Corpus of American Nordic Speech	(2015) 244 000 words. Speech. American Norwegian/Swedish. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Search the corpus
Corpus of Doctor-Patient Consultations from Ahus	(2015) 950 000 words. Speech. Transcriptions without audio files. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Get username and password - Search the corpus
The Lexicographic Corpus for Norwegian Bokmål	(2013) 100 mill words. Written text. Norwegian Bokmål. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Get username and password - Search the corpus
Nordic Dialect Corpus	(2013) 3 mill words. Speech. Nordic dialects. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Search the corpus
Nordic Syntax Database	(2013) 924 sentence judgments by Nordic dialect speakers. Accessible through interface. Licence: PUB BY-NC-SA. Licence conditions. - Download metadata - Search the database
The NORINT Corpus	(2017) Speech (110 000 words) and written text (53 000 words). Norwegian as second language. Accessible through interface. Licence: ACA. Licence conditions. - Download metadata - Search the corpus


Nasjonalbiblioteket
Språkbankens ressurskatalog BETA
Skriv inn søkeord
Søk
Nullstill
Om katalogen
Viser ressursene: 1 til 12 av totalt 92
Forrige
Neste
Språkbanken
CLARINO
Tekst
Parallel Corpus of documents from the Technical Regulations Information System for German-Spanish
PUB
17.03.2016
Tekst
Taggede bokmålstekster fra NBdigital (fritt tilgjengelige tekster)
PUB
07.03.2016
Tale
NB Tale - en grunnleggende akustisk-fonetisk taledatabase for norsk
PUB
25.02.2016
Tekst
N-grammer fra NBdigital
PUB
24.02.2016
Verktøy
NB N-gram
PUB
24.02.2016
Leksikon
Norsk Ordvev - Bokmål
PUB
22.02.2016
Tekst
Proceedings of Norwegian parliamentary debates (2008-2015)
PUB
17.02.2016
Leksikon
Norwegian-Vietnamese Dictionary
PUB
15.02.2016

CLARIN mission

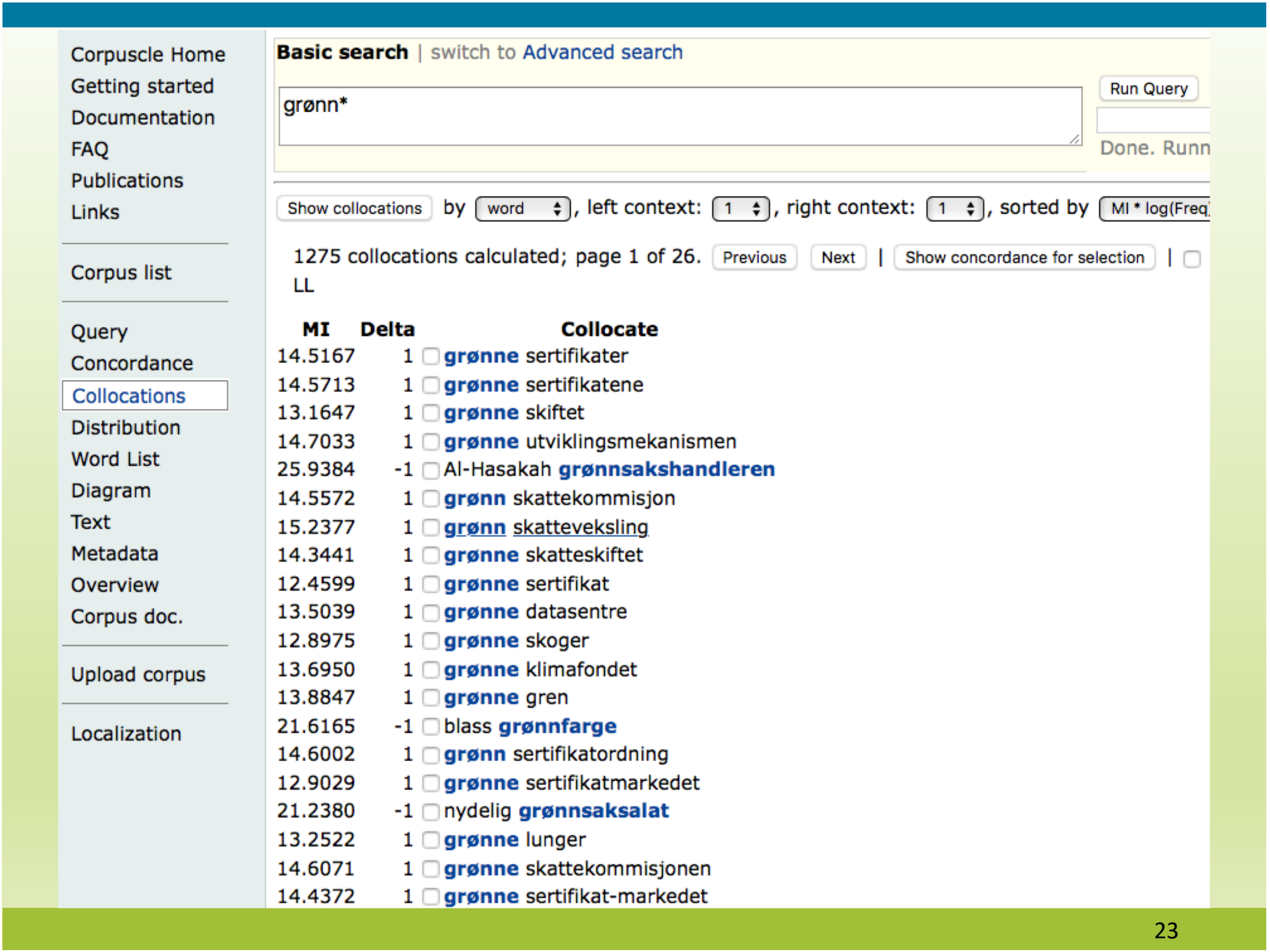
Not forcing a model on the DH or institutionalizing it, but contributing an infrastructure and meeting ground which aims to make

> “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)

Data is not enough

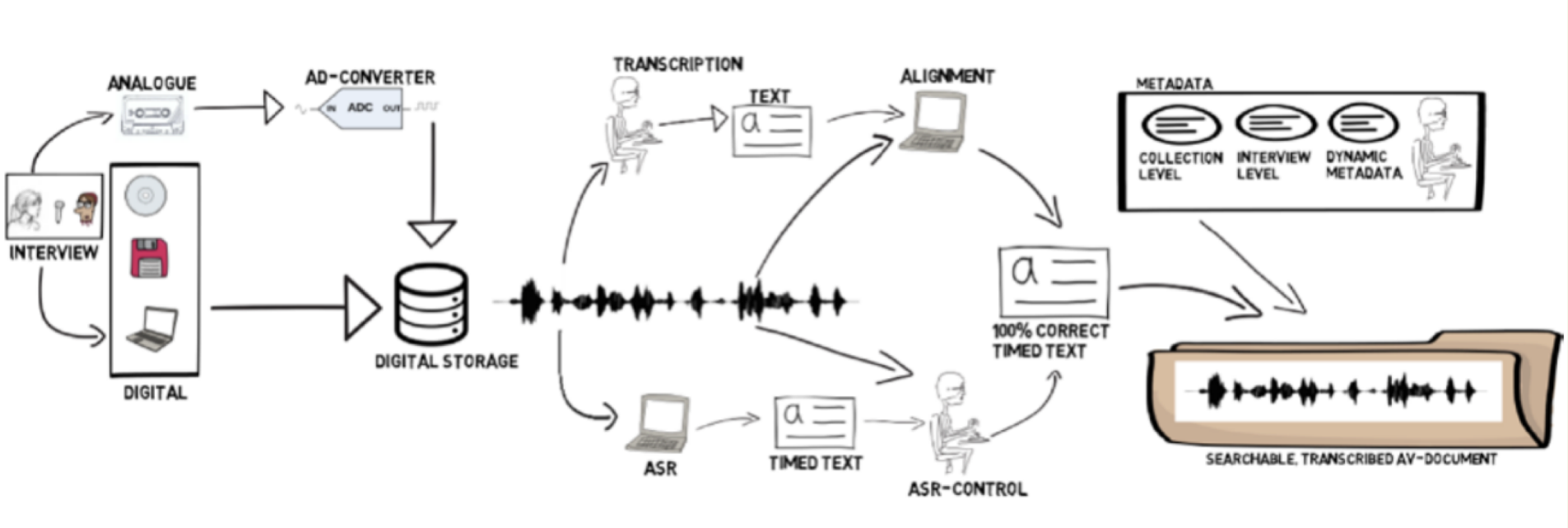
- Search in data (corpus search, treebank search, federated content search)
- Analyze data (annotate, parse, count)
- Visualize data (structures, relations, distributions)
- User support (knowledge centers, helpdesk)

Good tools are community-dependent and can only be designed, implemented, supported and maintained with the help of experts in the scientific field.


Corpuscle Home
Getting started
Documentation
FAQ
Publications
Links
Corpus list
Query
Concordance
Collocations
Distribution
Word List
Diagram
Text
Metadata
Overview
Corpus doc.
Upload corpus
Localization
Basic search | switch to Advanced search
grønn*
Run Query
Done. Runn
Show collocations by word, left context: 1, right context: 1, sorted by MI * log(Freq)
1275 collocations calculated; page 1 of 26. Previous Next | Show concordance for selection |
LL
MI Delta Collocate
14.5167 1 grønne sertifikater
14.5713 1 grønne sertifikatene
13.1647 1 grønne skiftet
14.7033 1 grønne utviklingsmekanismen
25.9384 -1 Al-Hasakah grønnsakshandleren
14.5572 1 grønn skattekommisjon
15.2377 1 grønn skatteveksling
14.3441 1 grønne skatteskiftet
12.4599 1 grønne sertifikat
13.5039 1 grønne datasentre
12.8975 1 grønne skoger
13.6950 1 grønne klimafondet
13.8847 1 grønne gren
21.6165 -1 blass grønnfarge
14.6002 1 grønn sertifikatordning
12.9029 1 grønne sertifikatmarkedet
21.2380 -1 nydelig grønnsaksalat
13.2522 1 grønne lunger
14.6071 1 grønne skattekommisjonen
14.4372 1 grønne sertifikat-markedet

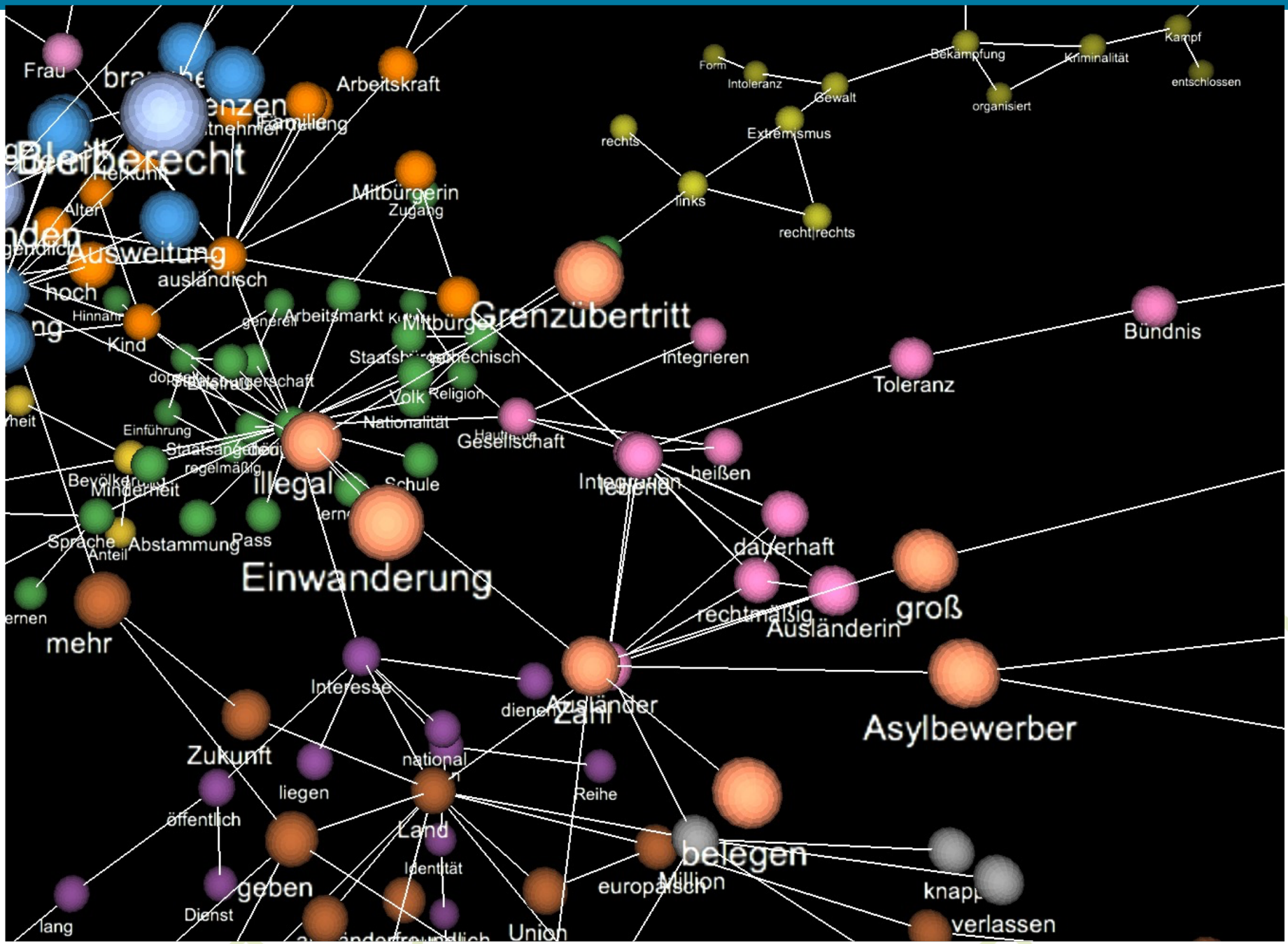
Research questions are the driving forces

- 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

- 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, 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

"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)

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

Virtual Language Observatory Search Help CLARIN

VLO / Faceted search / Search results

newspaper

Showing 1 to 10 of 21 results within selection for newspaper × Norwegian × Results per page: 10

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

Language

Type to filter or search for more

Norwegian ✖

German (696737)
French (64097)
Letzeburgesch; Luxembourgish (60302)
Slovenian (48521)
Finnish (24220)
Polish (12819)
Czech (5894)
Swedish (5889)
English (2517)
Croatian (2268)
more...

<< < 1 2 3 > >>

The Norwegian **Newspaper** Corpus

(Part of LRT + Open Submissions Data & Tools)

⊞ Dynamic, web-based **newspaper** corpus; 700 000 000 ws and growing; multitagged

NorGramBank **Newspaper** text (years 2012, 2013) in Norwegian Bokmål from the Norwegian **Newspaper** Corpus

(Part of Clarino Bergen Centre - INESS)

⊞ The "NorGramBank – **Newspaper** text (years 2012, 2013) in Norwegian Bokmål from the Norwegian **Newspaper** Corpus" treebank is a syntactically annotated corpus based on data taken from the years 2012 and 2013 from the Norwegian **Newspaper** Corpus (NCC). This treebank is part of INESS NorGramBank collection (see URL in metadat…

Norwegian **Newspaper** Corpus Bokmål

(Part of CLARINO UiB - Corpuscle)

⊞ The Norwegian **Newspaper** Corpus (NNC) Bokmål version is a large monitor corpus representing contemporary Norwegian language in the written variety Norwegian Bokmål. A corresponding corpus is available for Norwegian nynorsk, see URL in metadata. The corpus is compiled through daily harvesting

Conclusion and final remarks

- Making data ‘open’ is good but not enough; data needs to be FAIR
- Community-specific and data-specific infrastructures with advanced tools are needed
- European infrastructures scale up the amount and value of data
- CLARIN has a wide scope: languages, disciplines, countries, cultures, historical time spans
- Especially in the Humanities, infrastructures profit from collaboration with the GLAM sector (Galleries, Libraries, Archives, Museums)


CLARIN
http://clarin.eu