UiT

NOREGS
ARKTISKE
UNIVERSITET

TROLLing – eit ope arkiv for språkdata

17. møte om norsk språk (Mons 17)

UiB og HVL, Solstrand, Os

22.-24. november 2017

Philipp Conzett

Universitetet i Tromsø – Noregs arktiske universitet (UiT)

Disposisjon

1. Kva er TROLLing?
2. Arbeidsflyt og funksjonalitet i TROLLing
3. Kvifor skal forskarar arkivera dataa sine (i TROLLing)?
4. Samarbeid om språkdatahandtering

Kva er TROLLing?

Bakgrunn

- Idéen kom frå språkmiljøet på UiT. Behov for å kunna dela språkdata.
- Tok kontakt med UB UiT hausten 2013. UB hadde erfaring frå arbeid med open tilgang («open access») til forskingspublikasjonar.
- Arbeidsgruppe med språkvitarar, fagansvarlege for språkvitskap, rådgjevarar for open tilgang og IT-systemutviklarar.
- Rådgjevande gruppe med tre lingvistar frå Storbritannia og USA
- Lansering av TROLLing 18. juni 2014

opendata.uit.no


Foto: Kim G. Skytte

Innhald: dokumenttypar

Ope arkiv for språkdata og statistisk kode som kan delast

- Rekneark med annoterte data
- Tabellar og figurar
- Lydfiler
- Videoopptak
- Statistisk kode
- ...

	A	B	C	D	E	
1	word	first part	syntactic category of first part	second part	semantic category of second part	Bokmå
2	adelsløve	adel	N	løve	animal	nei
3	anstaltmaker	anstalt	N	maker	human	ja
4	apefant	ape	V (N?)	fant	human	nei
5	apegauk	ap/e	V (N?)			
6	apekatt	ape	N			
7	arbeidshest	arbeid	N			
8	arbeidsjern	arbeid	N			
9	arbeidsmaur	arbeid	N			
10	armodskrek	armod	N			
11	armodskryp	armod	N			

Nesset, Tore, 2014, "Replication data for: Metafor og metonymi: personkarakteriserende sammensatte substantiv i norsk", hdl:10037.1/10044, DataverseNO, V1; negative word formation Norw database 4 CSV version.csv [fileName]

Innhald: språk

- Slavisk: gamal austslavisk, gamal kyrkjeslavisk, mellomrussisk, olbansk/russisk, rumensk, russisk, tsjekkisk, ukrainsk
- Baltisk: latvisk
- Gresk: klassisk gresk
- Italisk: fransk, latin, rumensk
- Germansk: engelsk, islandsk, norsk, tidleg nyhøgtysk, tysk
- Andre: kinesisk, koreansk, nizaa, nordsamisk

Innhald: fagdisiplinar og emne

- Fagdisiplinar:
  fonetikk, fonologi, morfologi, semantikk, syntaks

- Tilnærmingar:
  synkroni, diakroni, korpuslingvistikk, språktileigning

- Emne:
  argumentstruktur, aspekt, genustilordning, kasus, konstruksjonsgrammatikk, metafor, objektsplassering, ordlaging, tospråklegheit, trykk, vokalar, ...

Statistikk per 18. november 2017

Datasett:		
Tal på datasett	60	I snitt per datasett:
Tal på datafiler	560	9
Tal på nedlastingar	3 503	58

Statistikk per 18. november 2017

Datasett:		
Tal på datasett	60	I snitt per datasett:
Tal på datafiler	560	9
Tal på nedlastingar	3 503	58

Forskarar (unike):	
Totalt	55
Frå Noreg	28
Frå UiT	23
Frå utlandet	27

Statistikk per 18. november 2017

Datasett:		
Tal på datasett	60	I snitt per datasett:
Tal på datafiler	560	9
Tal på nedlastingar	3 503	58

Forskarar (unike):	
Totalt	55
Frå Noreg	28
Frå UiT	23
Frå utlandet	27

Institusjonar (unike):	
Total	24
Frankrike	2
Kina	1
Latvia	2
Noreg	3
Russland	2
Storbritannia	3
Sør-Korea	2
Tsjekkia	1
Tyskland	4
USA	4

Statistikk per 18. november 2017

Datasett:		
Tal på datasett	60	I snitt per datasett:
Tal på datafiler	560	9
Tal på nedlastingar	3 503	58

Forskarar (unike):	
Totalt	55
Frå Noreg	28
Frå UiT	23
Frå utlandet	27

Ti-på-topp-datasett

Nedlastingar	**Språk**
329	Russisk
271	Russisk
215	Russisk
200	Russisk
200	Russisk
171	Norsk
148	Gamalkyrkjeslavisk
130	Gamalaustslavisk
123	Gamalkyrkjeslavisk
103	Spansk

Institusjonar (unike):	
Total	24
Frankrike	2
Kina	1
Latvia	2
Noreg	3
Russland	2
Storbritannia	3
Sør-Korea	2
Tsjekkia	1
Tyskland	4
USA	4

Arbeidsflyt og funksjonalitet i TROLLing

Tilrettelegging for gjenbruk

Tommelfingerregel: Ein fagfelle skal kunna gjenbruka datasettet også mange år etter at det er blitt publisert.

Følg retningslinene våre. I korte trekk:

- Bruk konsistente og forståelege filnamn.
- Lagre datasettet i arkivverdig/varig filformat i tillegg til originalformat.
- Beskriv datasettet i ei ReadMe-fil.

Meir info: https://site.uit.no/trolling/getting-started/

Arkivering: obligatoriske metadata

Title	Genusvariasjon i norsk skriftspråk
Author	Conzett, Philipp (UiT The Arctic University of Norway) - ORCID: 0000-0002
Contact	Use email button above to contact. Conzett, Philipp (UiT The Arctic University of Norway)
Description	Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva g genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrun undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" ne forklåra i ReadMe-fila. The data will be made openly available after the resu recognised scientific journal, but no later than January 1 2019. (2017-01-10
Subject	Arts and Humanities
Keyword	Norwegian Bokmål Nynorsk grammatical gender variation corpus linguistics

Arkivering: obligatoriske metadata

Title	Genusvariasjon i norsk skriftspråk
Author	Conzett, Philipp (UiT The Arctic University of Norway) - ORCID: 0000-0002
Contact	Use email button above to contact. Conzett, Philipp (UiT The Arctic University of Norway)
Description	Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva g genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrun undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" ne forklåra i ReadMe-fila. The data will be made openly available after the resu recognised scientific journal, but no later than January 1 2019. (2017-01-10
Subject	Arts and Humanities
Keyword	Norwegian Bokmål Nynorsk grammatical gender variation corpus linguistics

Arkivering: obligatoriske metadata

Title	Genusvariasjon i norsk skriftspråk
Author	Conzett, Philipp (UiT The Arctic University of Norway) - ORCID: 0000-0002
Contact	Use email button above to contact. Conzett, Philipp (UiT The Arctic University of Norway)
Description	Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva g genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrun undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" ne forklåra i ReadMe-fila. The data will be made openly available after the resu recognised scientific journal, but no later than January 1 2019. (2017-01-10
Subject	Arts and Humanities
Keyword	Norwegian Bokmål Nynorsk grammatical gender variation corpus linguistics

Arkivering: obligatoriske metadata

Title	Genusvariasjon i norsk skriftspråk
Author	Conzett, Philipp (UiT The Arctic University of Norway) - ORCID: 0000-0002
Contact	Use email button above to contact. Conzett, Philipp (UiT The Arctic University of Norway)
Description	Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva g genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrun undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" ne forklåra i ReadMe-fila. The data will be made openly available after the resu recognised scientific journal, but no later than January 1 2019. (2017-01-10
Subject	Arts and Humanities
Keyword	Norwegian Bokmål Nynorsk grammatical gender variation corpus linguistics

Arkivering: obligatoriske metadata

Title	Genusvariasjon i norsk skriftspråk
Author	Conzett, Philipp (UiT The Arctic University of Norway) - ORCID: 0000-0002
Contact	Use email button above to contact. Conzett, Philipp (UiT The Arctic University of Norway)
Description	Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva g genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrun undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" ne forklåra i ReadMe-fila. The data will be made openly available after the resu recognised scientific journal, but no later than January 1 2019. (2017-01-10
Subject	Arts and Humanities
Keyword	Norwegian Bokmål Nynorsk grammatical gender variation corpus linguistics

Arkivering: obligatoriske metadata

Title	Genusvariasjon i norsk skriftspråk
Author	Conzett, Philipp (UiT The Arctic University of Norway) - ORCID: 0000-0002
Contact	Use email button above to contact. Conzett, Philipp (UiT The Arctic University of Norway)
Description	Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva g genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrun undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" ne forklåra i ReadMe-fila. The data will be made openly available after the resu recognised scientific journal, but no later than January 1 2019. (2017-01-10
Subject	Arts and Humanities
Keyword	Norwegian Bokmål Nynorsk grammatical gender variation corpus linguistics

Arkivering: valfrie metadata

- Finansiør
- Relatert publikasjon
- Innsamlingsperiode
- Innsamlingsmetode
- Undersøkt tidsperiode
- Datakjelder
- Geografisk informasjon: Kva område dekkjer undersøkinga
- ...

Filopplasting (+ ev. embargo)

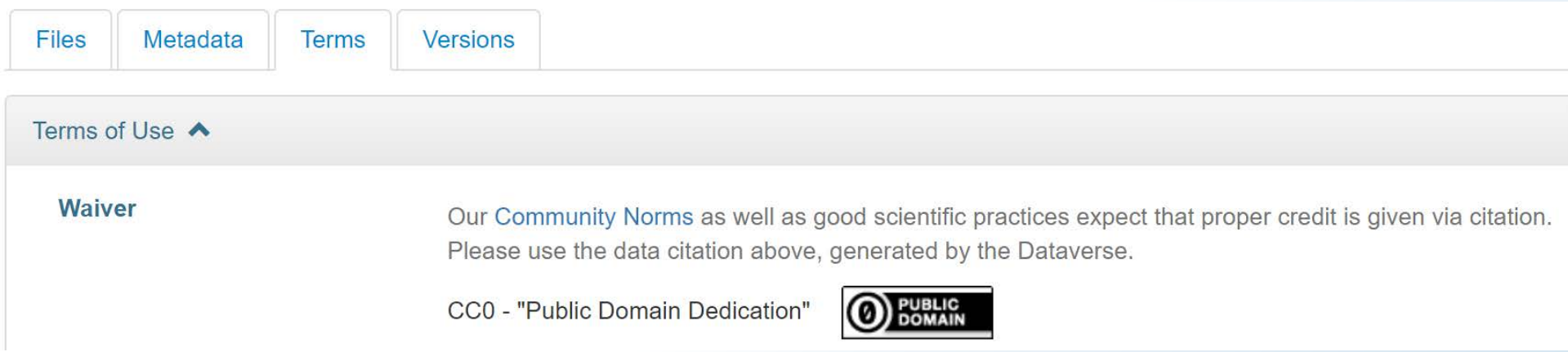
- Mogleg å leggja embargo på filtilgang:


Dette datasettet inneheld materialet ifrå ei undersøking av genusvariasjon i norsk skriftspråk. Undersøkinga har sitt utspring i eit oppdrag eg fekk ifrå Språkrådet om å kartleggja i kva grad valfridomen i genusnormeringa i bokmål og nynorsk er nytta i skriftlege kjelder. Datagrunnlaget og metoden som er nytta i undersøkinga, er beskrivne i prosjektrapporten (sjå "Related Publication" nedanfor). Oppsettet på filene er forklåra i ReadMe-fila. The data will be made openly available after the results have been published in a recognised scientific journal, but no later than January 1 2019.

Bokmalsmateriale_m+n.ods

Lisensar: vilkår for gjenbruk

- Standardlisens i TROLLing: Creative Common / CC0
  = ingen restriksjonar på gjenbruk
- Men forventar kreditering gjennom tilvising



- Mogleg å bruka andre lisensar

Kuratering

- Fagansvarleg på UB UiT sjekkar om datasettet er tilrettelagt for gjenbruk.

Publisering: Referanse med varig URL

Holliday, Jeff; Turnbull, Rory; Eychenne, Julien, 2016, "K-SPAN (Korean Surface Phones and Neighborhoods)", doi:10.18710/TWM79F, DataverseNO, V1, UNF:6:yYK7L4GaX1TTyBwMMhwZbg==

Versjonering: V1 → V2

Holliday, Jeff; Turnbull, Rory; Eychenne, Julien, 2016, "K-SPAN (Korean Surface Phones and Neighborhoods)", doi:10.18710/TWM79F, DataverseNO V1
UNF:6:yYK7L4GaX1TTyBwMMhwZbg==

Holliday, Jeffrey J.; Turnbull, Rory; Eychenne, Julien, 2016, "K-SPAN (Korean Surface Phones and Neighborhoods)", doi:10.18710/TWM79F, DataverseNO V2
UNF:6:NWbRmiBvO5wWcDN2QHCQJw==

Versjonskontroll

- Oversikt over endringar mellom versjonar:

Files | Metadata | Terms | Versions

View Differences

	Dataset	Summary	Contributors	Published
☐	2.0	**Files (Added: 3; Removed: 3);** View Details	Julien Eychenne, Philipp Conzett	September 1, 2016
☐	1.1	**Citation Metadata:** Author (3 Changed); View Details	Julien Eychenne, Philipp Conzett	June 3, 2016
☐	1.0	This is the first published version.	Julien Eychenne, Philipp Conzett	June 3, 2016

Tilvising i publikasjonar: tidsskriftsartikkel

DE GRUYTER MOUTON Folia Linguistica 2016; 50(2): 385–412

Fabian Barteld, Stefan Hartmann and Renata Szczepaniak*

The usage and spread of sentence-internal capitalization in Early New High German: A multifactorial approach

. . .

Resources

The dataset and R script have been made available at the Tromsø Repository for Language and Linguistics (TroLLing): http://dx.doi.org/10.18710/SJ4OQE.

(https://stefanhartmanneu.files.wordpress.com/2017/05/barteld-et-al-2016.pdf)

Tilvising i publikasjonar: ph.d.-avhandling

In the beginning was the word

A study of monolingual
and bilingual children's lexicons

Pernille Hansen

Center for Multilingualism in Society across the Lifespan
Department of Linguistics and Scandinavian Studies
Faculty of Humanities
University of Oslo

Dissertation submitted for the degree of PhD
December 2016

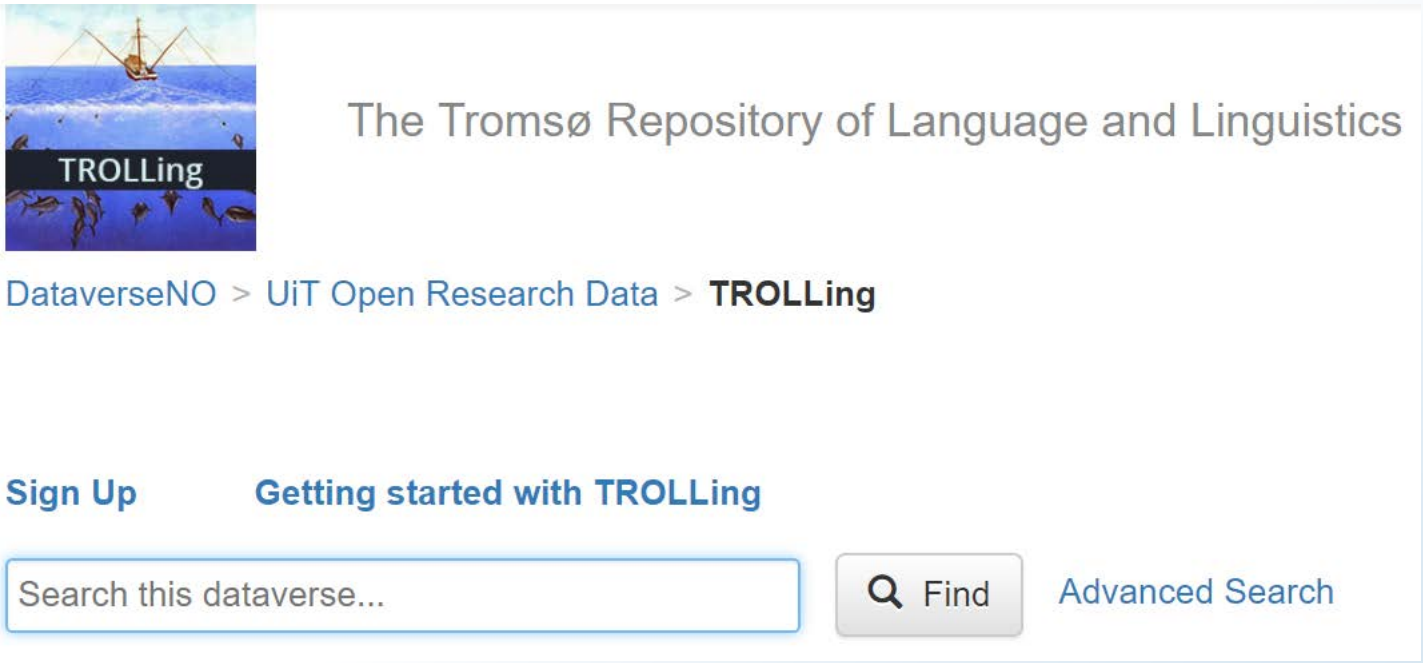
. . .

Hansen, P. (2016). Replication data for: What makes a word easy to acquire? The effects of word class, frequency, imageability and phonological neighbourhood density on lexical development. UiT Open Research Data Dataverse, V1. doi:10.18710/JEWIVW

(http://www.hf.uio.no/multiling/personer/postdoktorer/perniha/dissertation-pernille-hansen-without-papers.pdf)

Gjenbruk: søk i TROLLing

opendata.uit.no

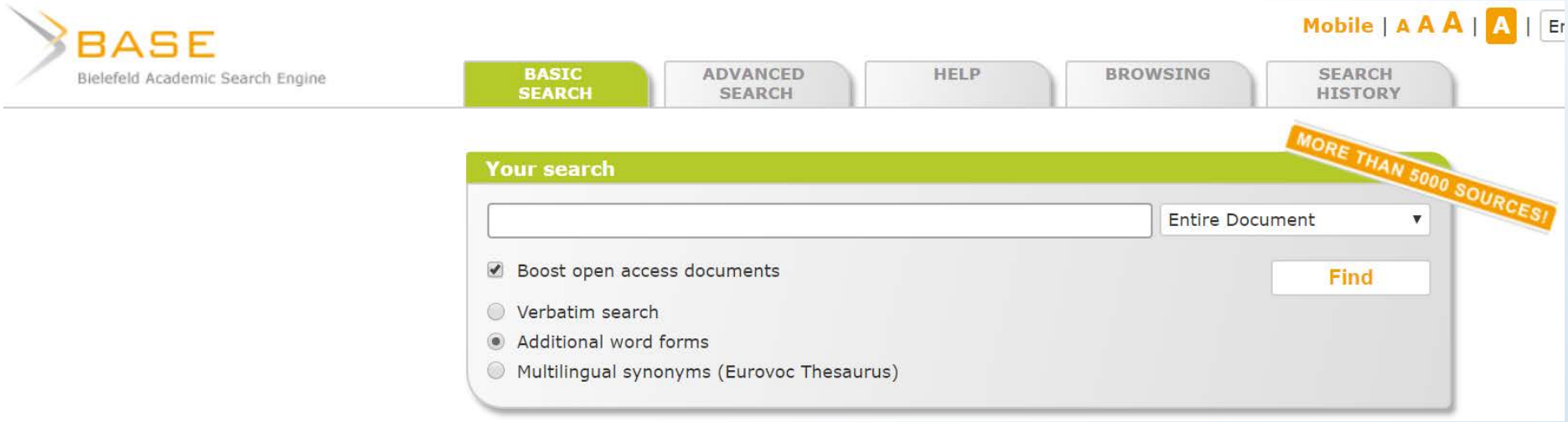


Gjenbruk: søk i generelle søkjetenester: DataCite

https://search.datacite.org/

Gjenbruk: søk i generelle søkjetenester: BASE

https://www.base-search.net/

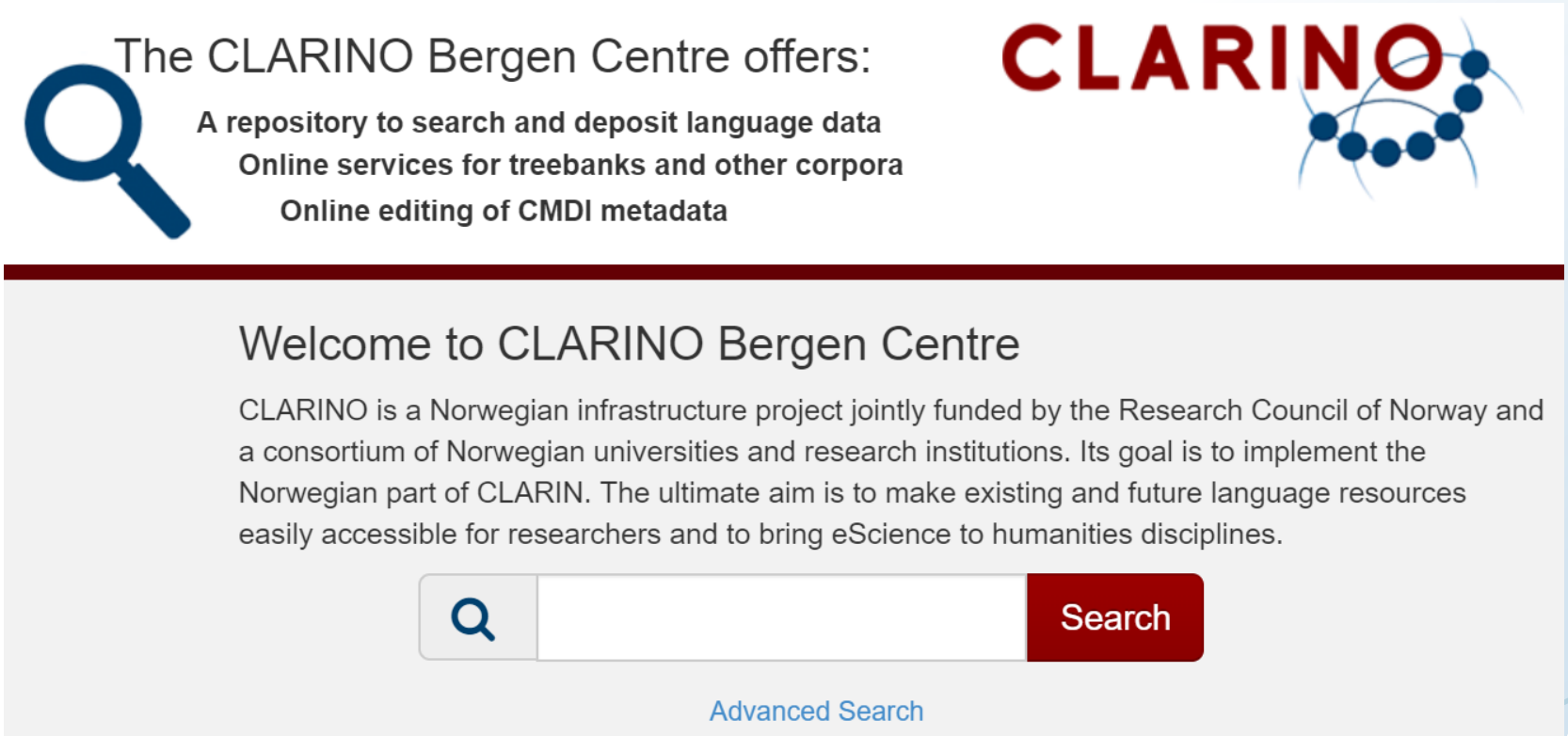


Gjenbruk: søk i generelle søkjetenester: Oria

oria.no

Gjenbruk: søk i CLARINO

https://repo.clarino.uib.no/



Kvifor dela data (i TROLLing)?

Det kortet svaret: Fordi du (snart) må!

Krav frå finansiørar: EU

Frå 2017:

European Commission

Horizon 2020
European Union funding
for Research & Innovation

«By extending the pilot, open access becomes the default setting for research data generated in Horizon 2020.»

«... as open as possible, as closed as necessary ...»

(Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020)

Krav frå finansiørar: NFR

Frå 2017:

«Forskningsrådets policy følger "åpen som standard"-prinsippet når det gjelder tilgang til forskningsdata.»

(Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd)

Forventingar frå tidsskrift: *Language* (LSA)

LANGUAGE

A JOURNAL OF THE LINGUISTIC SOCIETY OF AMERICA

«Where the data is not publicly available, (and there are often good reasons why it is not), authors should explain why it is not made available ...»

(Notes to Contributors to *Language*, published by the Linguistic Society of America)

Krav frå tidsskrift: *Poljarnyj vestnik* (UiT)

Poljarnyj vestnik
Norwegian Journal of Slavic Studies

«Poljarnyj vestnik acknowledges the need for Slavic linguists to archive their data in a safe place – and to share the data with their colleagues. Authors of articles for Poljarnyj vestnik are therefore requested to archive their data and code at TROLLing (the Tromsø Repository of Language and Linguistics).»

(Author Guidelines for *Poljarnyj vestnik*)

Krav frå institusjonar: UiT

«Pkt. 4.4 Forskeren skal gjøre forskningsdata åpent tilgjengelig for videre bruk for alle relevante brukere, så fremt det ikke er juridiske, etiske, sikkerhetsmessige eller kommersielle grunner til ikke å gjøre det.»

(Prinsipper og retningslinjer for forvaltning av forskningsdata ved UiT. Vedtekne av Universitetsstyret 9. mars 2017. Gjeldande frå 1. september 2017.)

... og fleire vil følgja etter

- Kunnskapsdepartementet kjem med nasjonal strategi for tilgjengeleggjering og deling av forskingsdata innan utgangen av 2017.

- Klare føringar for UH-sektoren

Det litt lengre svaret: Fordi forskinga tener på det!

Tilgjengeleggjering og deling av forskingsdata bidreg til

- «Forbedret kvalitet i forskningen gjennom bedre mulighet til å bygge på tidligere arbeider og sammenstille data på nye måter»

(Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd)

Det litt lengre svaret: Fordi forskinga tener på det!

Tilgjengeleggjering og deling av forskingsdata bidreg til

- «Forbedret kvalitet i forskningen gjennom bedre mulighet til å bygge på tidligere arbeider og sammenstille data på nye måter»
- **«Gjennomsiktighet i forskingsprosessen og bedre mulighet for etterprøvbarhet av vitenskapelige resultater»**

(Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd)

Det litt lengre svaret: Fordi forskinga tener på det!

Tilgjengeleggjering og deling av forskingsdata bidreg til

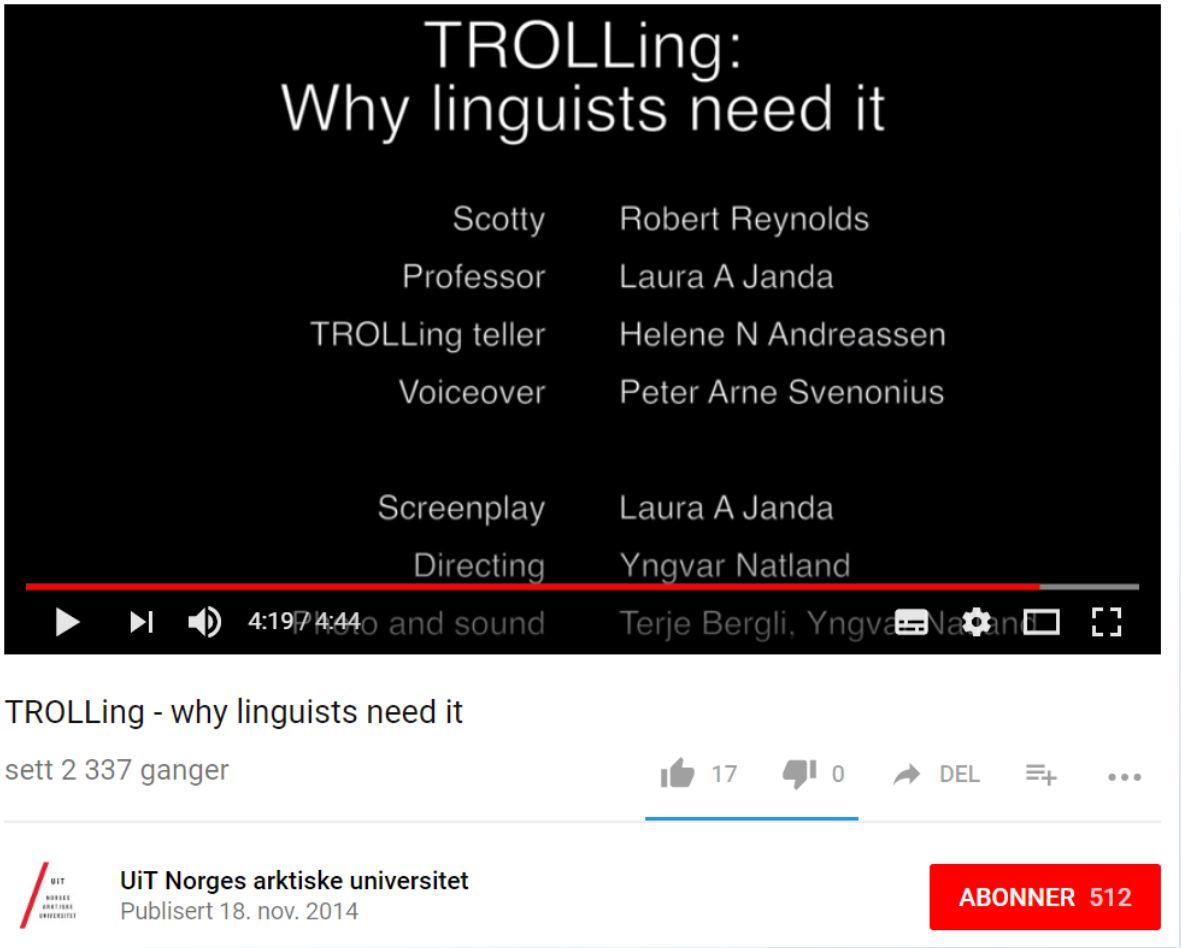
- «Forbedret kvalitet i forskningen gjennom bedre mulighet til å bygge på tidligere arbeider og sammenstille data på nye måter»
- «Gjennomsiktighet i forskingsprosessen og bedre mulighet for etterprøvbarhet av vitenskapelige resultater»
- **«Økt samarbeid og mindre duplisering av forskningsarbeid»**

(Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd)

Framleis ikkje overtydd?

YouTube NO

http://youtu.be/uEf0c0NT9_A



Samarbeid om språkdatahandtering: Kva?

- Kuratering: Korleis kan vi best mogleg leggja forskingsdata til rette for deling?

Samarbeid om språkdatahandtering: Kva?

- Kuratering: Korleis kan vi best mogleg leggja forskingsdata til rette deling?

- **Metadata: Korleis skal vi beskriva forskingsdata for at dei kan bli gjenfunne også på tvers av arkiv?**

Samarbeid om språkdatahandtering: Kva?

- Kuratering: Korleis kan vi best mogleg leggja forskingsdata til rette deling?

- Metadata: Korleis skal vi beskriva forskingsdata for at dei kan bli gjenfunne også på tvers av arkiv?

- **Tilvising: Korleis skal ein visa til forskingsdata i publikasjonar?**

Samarbeid om språkdatahandtering: Kva?

- Kuratering: Korleis kan vi best mogleg leggja forskingsdata til rette deling?
- Metadata: Korleis skal vi beskriva forskingsdata for at dei kan bli gjenfunne også på tvers av arkiv?
- Tilvising: Korleis skal ein visa til forskingsdata publikasjonar?
- **Opplæring og informasjon: Korleis kan vi nå ut til forskarane og andre aktørar?**

Samarbeid om språkdatahandtering: Kven?

- Arkivtenester og forskingsinfrastrukturnettverk:
  - CLARIN(O)
  - NSD
  - Språkbanken (Nasjonalbiblioteket)
  - TROLLing
  - ...
- Internasjonale organisasjonar og nettverk
  - Research Data Alliance: Interessegruppe for språkvitskap («Linguistics Data IG»)
  - ...
- Språkrådet
- Forlag og tidsskrift
- **Forskarane**
- ...


Takk for merksemda!
NB! Hugs TROLLing-plakaten!
opendata.uit.no

Referansar

Author Guidelines for *Poljarnyj vestnik*. Tilgjengeleg her.

Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Tilgjengeleg her.

Notes to Contributors to *Language*, published by the Linguistic Society of America. Tilgjengeleg her.

Prinsipper og retningslinjer for forvaltning av forskningsdata ved UiT. Tilgjengeleg her.

Tilgjengeliggjøring av forskningsdata - Policy for Norges forskningsråd. Tilgjengeleg her.