UiT Noregs arktiske universitet

Kor FAIR er norske språkdata?

Fagleg seminar for språksamlingane
Universitetet i Bergen -- digitalt
19.-20. november 2020

Philipp Conzett
UiT Noregs arktiske universitet
ORCID: https://orcid.org/0000-0002-6754-7911

Plan for presentasjonen

1 Kva er FAIR?
2 Korfor treng vi FAIR?
3 Kor FAIR er norske språkdata*?
4 Spørsmål og diskusjon

(*Ordet *data* er her brukt som samlenemning for alle typar ressursar/materiale som kan brukast til språkforsking.)

1 Kva er FAIR?

Kva er FAIR?

- Eit sett med generelle prinsipp for god handtering og tilgjengeleggjering av forskingsdata

- Data som er FAIR,
  - kan gjenfinnast,
  - er tilgjengelege,
  - er interoperable, og
  - kan gjenbrukast.

På engelsk:

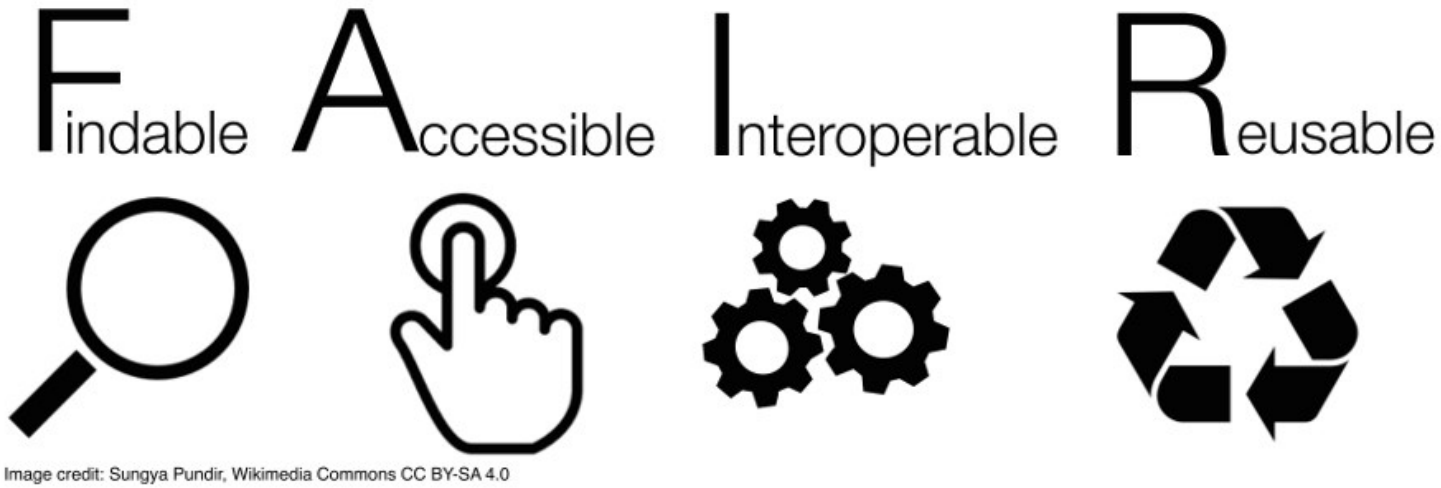
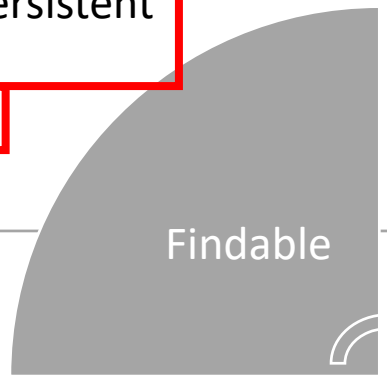


Image credit: Sungya Pundir, Wikimedia Commons CC BY-SA 4.0

FAIR-prinsippa: Findable

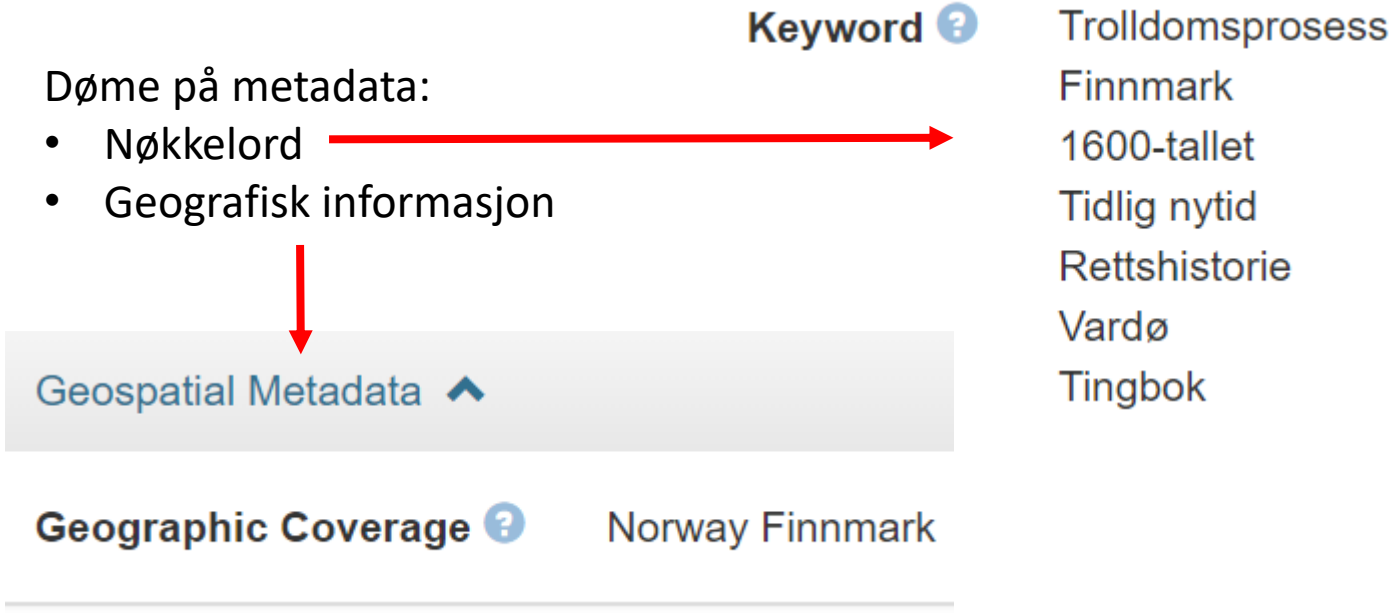
- Publisert med persistent identifikator
- Gode metadata
- Indeksert



Metadata = beskriving av data

Døme på metadata:

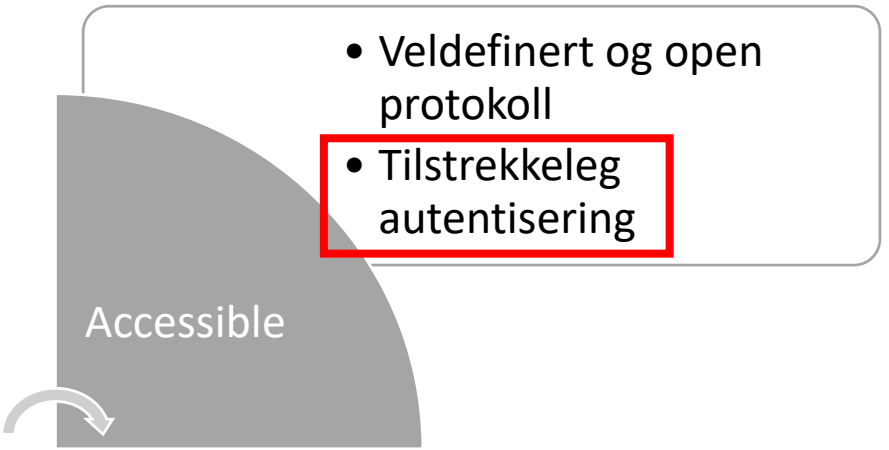
- Nøkkelord
- Geografisk informasjon



Hagen, Rune Blix, 2019, "Rettsforfulgte trollfolk i Finnmark, 1593-1692", https://doi.org/10.18710/OWP5IP, DataverseNO, V1

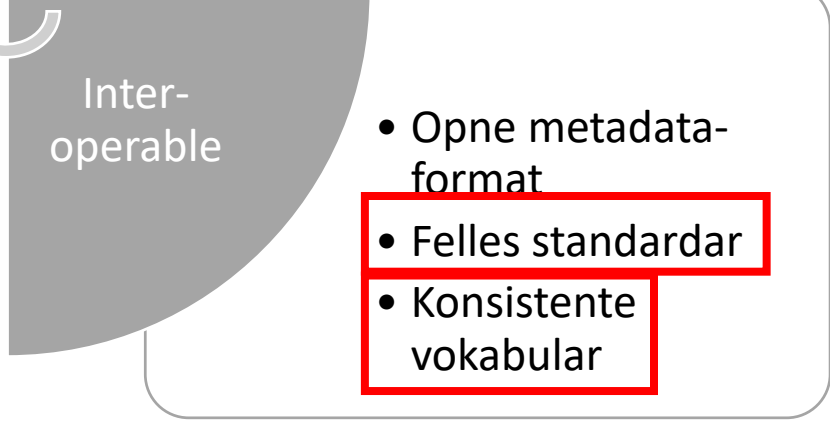
DOI = Digital Object Identifier = ein type persistent identifikator ~ varig lenkje/URL

FAIR-prinsippa: **A**ccessible



- *Accessible* har mest med tekniske aspekt ved dataarkiv å gjera. Men:
- Når ein publiserer dataa sine, bør ein velja eit arkiv som gjer dei tilgjengelege i tråd med innhaldet. Døme:
- Data som ikkje kan gjerast ope tilgjengelege, kan kanskje delast i eit arkiv der dei som ønskjer å lasta ned data, må registrera seg og logga inn. Då treng ein tilstrekkeleg autentisering.

FAIR-prinsippa: Interoperable

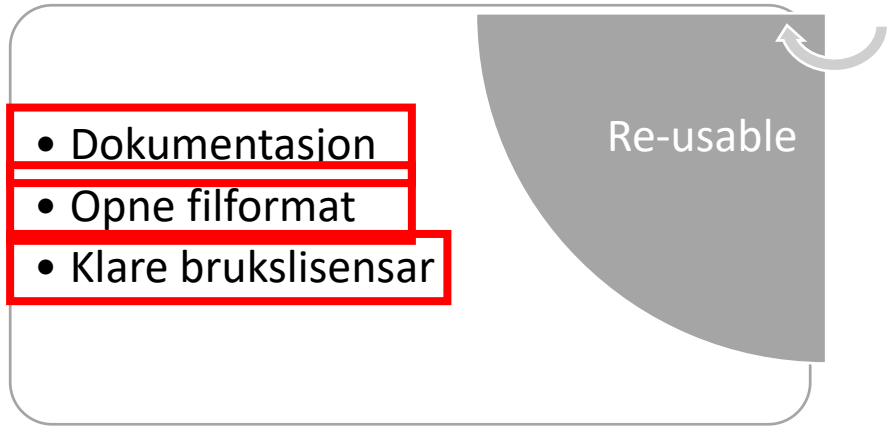


- Bruk felles metadatastandardar. Det gjeld både
  - generelle metadata, t.d. internasjonalt datoformat (t.d. ISO-8601): ÅÅÅÅ-MM-DD (2019-12-09), og
  - fagspesifikke metadata, t.d. Data Documentation Initiative (DDI) = internasjonal standard for beskriving av data brukte i spørjeskjema og andre observasjonsmetodar i samfunnsfag og helsefag.
- Bruk konsistente metadatavokabular, t.d. DDI-vokabularet for aggregeringsmetode (Aggregation Method); utdrag:
- Interoperabilitet mogleggjer søk og gjenbruk på tvers av datasett og arkiv.

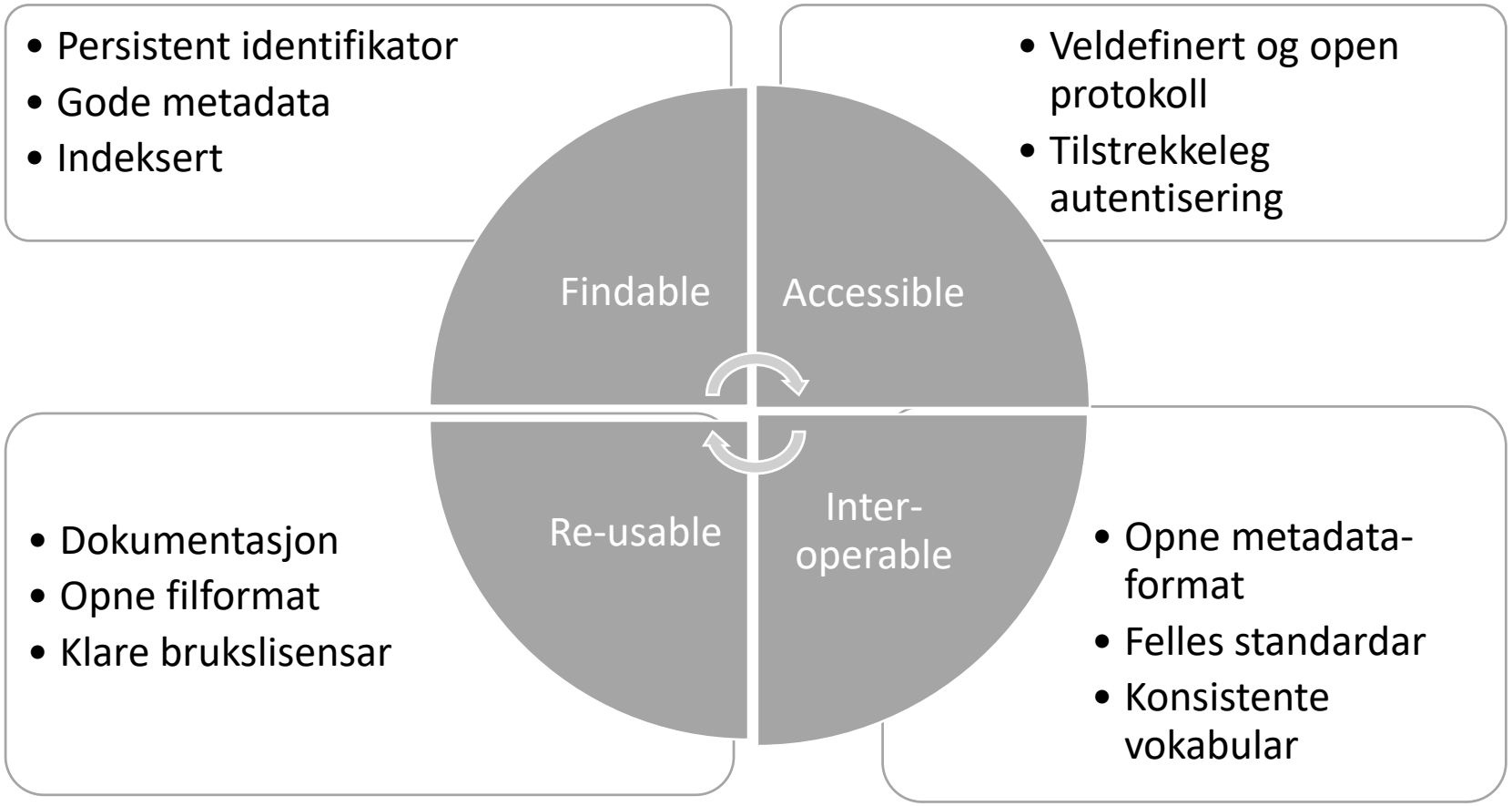
Value of the Code	Descriptive Term of the Code	Definition of the Code
Maximum	Maximum	The highest value attained or recorded.
Minimum	Minimum	The lowest value attained or recorded.

FAIR-prinsippa: **R**e-usable

- Dokumenter data, slik at dei er forståelege og kan gjenbrukast av fagfellar.
- Arkiver data i føretrekte/arkivverdige filformat slik at filene kan opnast og lesast på lang sikt, t.d. rein tekst (.txt) i tillegg til Excel (.xlsx).
- Definer ein klar brukslisens for dataa dine slik at dei som ønskjer å bruka dei, veit kva dei har lov til å gjera med dei. Døme: Creative Commons (CC)-lisensar



Til saman blir det FAIR:



2 Korfor treng vi FAIR?

Gode for forskingsdatahandtering

FAIR-prinsippa hjelper oss med å handtera å dela forskingsdataa våre på ein god måte, slik at

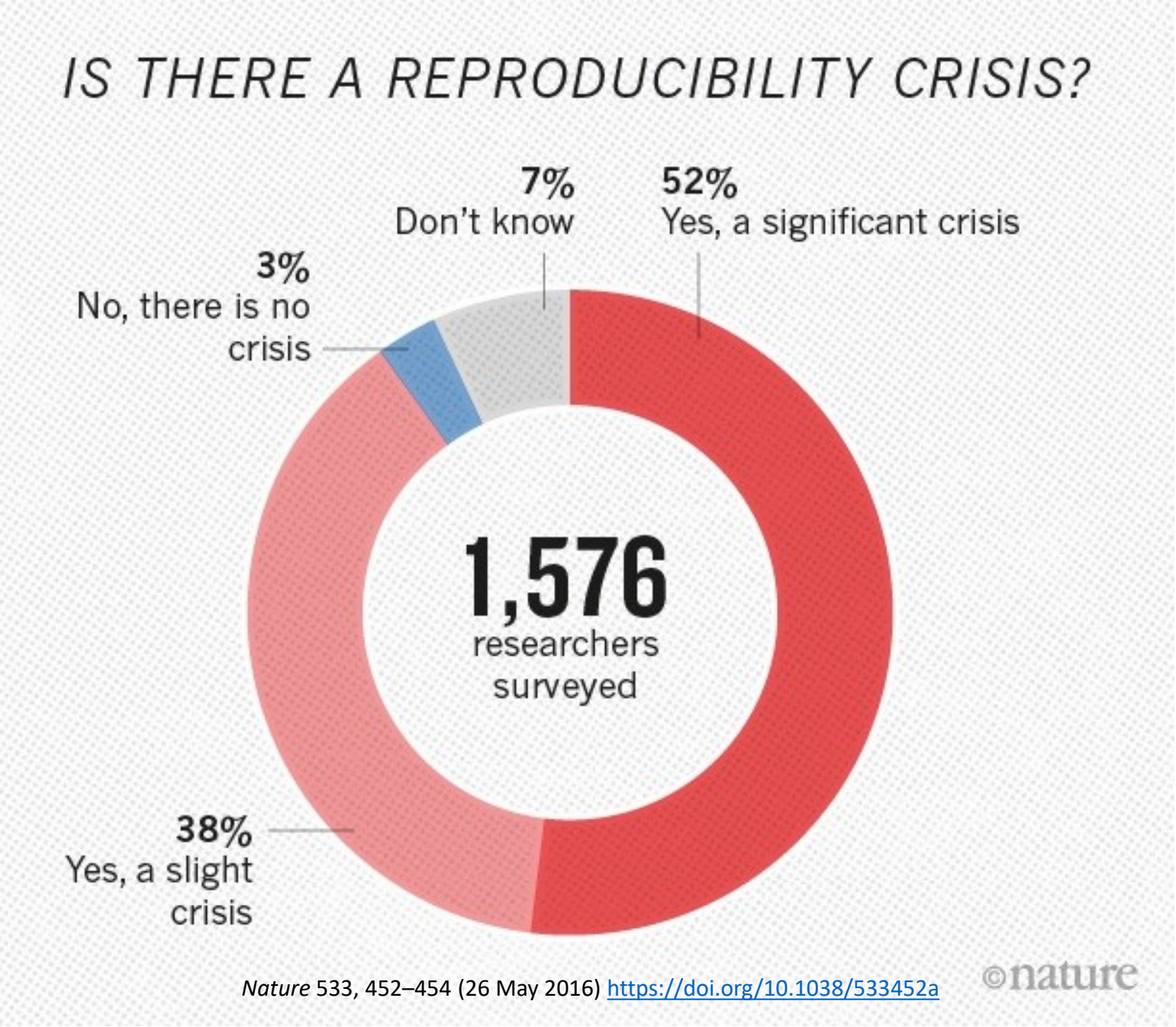
- for at forskinga vår i størst mogleg grad skal kunna etterprøvast og reproduserast*, og
- for at forskingsdata skal kunna gjenbrukast (på andre måtar enn til etterprøving og reproduksjon).

(*Skiljet mellom reproduserbarheit og replikerbarheit er diskutert i m.a. Plesser (2018).)

Etterprøvbarheit/reproduserbarheit

Det er vel ei sjølvfølgje at forskingsresultat som framstilte i publikasjonar skal vera etterprøvbare!?

>> Nei!



Nature 533, 452–454 (26 May 2016) https://doi.org/10.1038/533452a

Meir enn 50 % av dei spurde forskarane er samd i at forskinga er råka av ei **reproduserbarheits-krise**.

Meir enn 70 % av dei spurde forskarane har prøvd, men mislukkast med å reproducera forskingsresultata av ein **annan** forskar. Meir enn halvparten har mislukkast med å reproducera sine **eigne** resultat.

Kva er hovudgrunnen?
>> Selektiv rapportering, t.d. «rosinplukking» av data for å stø opp om ein hypotese

Kva har det med språkdata å gjera?

Undersøkinga som Nature refererer til, gjeld sikkert berre realfag, psykologi og slike fag, men ikkje språkvitskap!?

>> Nei!

Thomason 1994

Sally Thomason, i 1994 redaktør for tidsskriftet Language oppdaga ofte **problematiske aspekt ved datagrunnlaget** for artikkelmanus, «so frequently, in fact, that the assumption that the data in accepted papers is reliable began to look questionable». (Thomason 1994: 409)

Dømet er henta frå Berez-Kroeker et al. (2018: 8)

Gawne et al. 2017

«In a **survey of one hundred descriptive grammars** from a ten-year span between 2003 and 2012, Gawne and colleagues (2017) found that even with the benefit of years of pervasive discussion of data management methods in language documentation, **very few authors in this genre make their methods or data sources explicit in their writing**.» (Berez-Kroeker et al. 2018, 9) (mine framhevingar, PhC)

Berez-Kroeker et al. 2017

«In a **survey of 270 articles from nine top international linguistics journals** from the same time period, Berez-Kroeker and colleagues (2017a) found that **scant few journal authors met any – let alone all – of the survey's metrics for basic transparency of data and methodology**, including sufficient citation of numbered examples from unpublished sources, or a minimal description of methods of data collection and analysis.» (Berez-Kroeker et al. 2018, 9) (mine framhevingar, PhC)

3 Kor FAIR er norske språkdata? --- Eller …

… kva har dette med Noreg å gjera?

Funna som Thomason, Gwane et al. og Berez-Kroeker et al. viser til, gjeld sikkert berre utanlandsk språkforsking!?

>> Nei!

Ein skal ikkje hengja ut kollegaene sine, men …

Sollid, Conzett & Johansen 2014

Ei undersøking av genus og substantiv bøying i nordnorske kontaktvarietetar

--- Resultat:

Table 7. Distribution of indefinite articles according to location and traditional gender (numbering of locations refers to Figure 1)

Location	Indef art								
	ei (# = 113)			*en* (# = 998)			*et* (# = 264)		
	F	M	N	F	M	N	F	M	N
1 Karlsøy	22	1	0	0	48	0	0	0	23
2 Hammerfest	27	0	0	3	63	3	0	0	13
3 Kjøllefjord	23	0	1	4	68	1	0	0	18
4 Vardø	18	0	0	9	161	0	0	1	41
5 Kirkenes	16	0	0	9	100	0	0	1	41
6 Kvænangen	1	0	0	19	154	0	0	2	32
7 Lakselv	3	0	0	20	113	2	0	0	54
8 Tana	0	0	0	36	106	3	0	0	18
9 Kautokeino	0	1	0	21	55	0	0	0	20
Total: 1375	110	2	1	121	868	9	0	4	260

--- Metode:

4. The Nordic Dialect Corpus – methodological considerations

4.1 The Nordic Dialect Corpus and our corpus

This study uses data from the Nordic Dialect Corpus (NDC), a collection of speech data from Faroe Islands, Iceland, Norway, Sweden, Denmark and Finland. In the Norwegian part of the corpus, there are over 2 million words from 137 measure points. Our data comes from nine field sites: Karlsøy, Kvænangen, Hammerfest, Kjøllefjord, Kirkenes, Vardø, Lakselv, Tana and Kautokeino (cf. Figure 1).

--- Men kva med grunnlagsdataa?

References

Bull, T. [1990] 1996. Målet i Troms og Finnmark. In *Nordnorske dialektar*, E. H. Jahr & O. Skare (eds), 157–174. Oslo: Novus.

...

Sollid, H. 2005. *Språkdannelse og -stabilisering i møtet mellom kvensk og norsk*. Oslo: Novus.

Eller: Kor finn vi tabellar som denne?

NP_nr	NP	genus	trad_gen	tall	bestemthet	subst_bøying	ub_art	possessiv
70	pappa	m		sg	u	0		
72	en gang	m		sg	u	0	en	
73	uka	f		sg	b	a		
74	en ronn ting	m		sg	u	0	en	
75	båtn	m		sg	b	n		
76	fjorn	m		sg	b	n		
77	sjøen	m		sg	b	en		
80	søstra min	f		sg	b	a		X_min

…

NB! Sollid, Conzett & Johansen 2014 er på langt nær det «verste» eksempelet! Bør undersøkjast! --- ?Conzett 2021?

Her kjem eit (meir) eksemplarisk eksempel:

Flick 2020a & Flick 2020b

Ein studie av utviklinga av den bestemde artikkelen i gamalhøgtysk

1. Publikasjon med open tilgang:

language science press

Open Access

PDF

2. Grunnlagsdata tilgjengelege:

Alle Korpusdaten, Annotationsrichtlinien und R-Skripte wurden in Flick (2020) veröffentlicht, so dass das Vorgehen sowie die Ergebnisse der vorliegenden Untersuchung transparent dokumentiert sind und zukünftige Studien an die Materialien anknüpfen können.

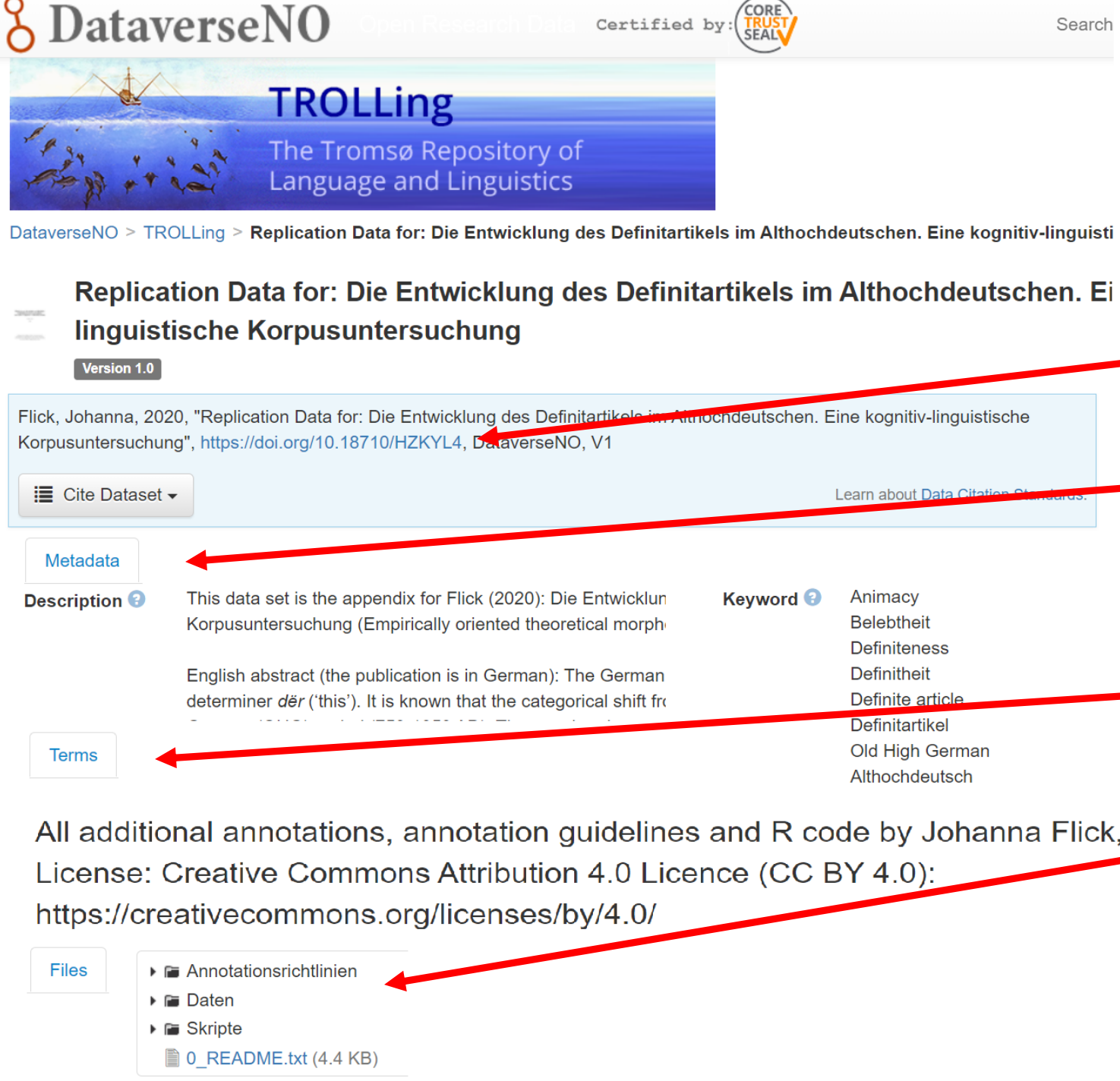
(Alle korpusdataa, annotasjonsretningslinene og R-skript blei publiserte i Flick (2020), slik at prosedyren og resultata frå denne studien er transparent dokumenterte og framtidige studiar kan byggja på desse materialane.)

Literaturverzeichnis

Abbott, Barbara. 2007. Definiteness and indefiniteness. In Laurence R. Horn & Gregory Ward (Hrsg.), *The handbook of pragmatics*, 3. Aufl. (Blackwell Handbooks in Linguistics 16), 122–149. Malden: Blackwell.

…

Flick, Johanna. 2020. *Replication Data for: Die Entwicklung des Definitartikels im Althochdeutschen. Eine kognitiv-linguistische Korpusuntersuchung*. Version 1. DOI:10.18710/HZKYL4



Viktige FAIR-element er på plass:

- DOI
- Metadata
- Brukslisens
- Tilgang til filene, inkl. dokumentasjon

Kva med kjeldedata?

- Til no har vi snakka mest om **resultatdata**, t.d. prosesserte og analyserte data som dannar grunnlaget for ein artikkel- eller bokpublikasjon.
- Men kva med **kjeldedata**? For å kunna referera til kjeldedata på ein god måte bør også dei vera så FAIR som moglege.

Men først: Kva treng vi til ein god referanse?

Andreassen et al. (2019): Tromsø recommendations for citation of research data in linguistics

Eit sett med tilrådingar for korleis ein bør referera til språkdata

«Language and linguistics datasets are often not cited, or cited imprecisely, because of confusion surrounding the proper methods for citing them.»

The Tromsø recommendations «propose components of data citation for referencing language data, both in the bibliography and in the text of linguistics publications».


Foto: Helene N. Andreassen

Mal for utvida bibliografisk referanse

The template for an **expanded bibliographic reference** to a dataset resource, including *conditional elements* (i.e. required in certain cases depending on resource characteristics) is:

Author, *Other Attribution (Roles),* **Date, Title, Publisher, Locator,** *Version*, *Date accessed.*

t.d. Collector, Consultant, …

t.d. DOI

Eit eksempel frå Språksamlingane

Norsk Ordboks nynorskkorpus

Manglar fleire grunnleggjande FAIR-element

The template for an **expanded bibliographic reference** to a dataset resource, including *conditional elements* (i.e. required in certain cases depending on resource characteristics) is:

Author, *Other Attribution (Roles),* **Date, Title, Publisher, Locator,** *Version, Date accessed.*

- **Author** → Eller opphavsmann/-institusjon?
- **Date** → Når er korpuset publisert?
- **Publisher** → Kven er utgjevaren?
- **Locator** → Har ein URL, men ingen persistent identifikator
- *Version* → Kva er den gjeldande versjonen?

I tillegg: Kva bruksvilkår / lisens er det som gjeld for korpuset?

… eller:

Kva får eg lov til å gjera med data som er henta ut av Norsk Ordboks nynorskkorpus?

Genus	Genus_kong	Genus_b	Aar	Venstrekontekst	Hovudord	Hoegrekontekst
m	m		2008	. I dagboka finn vi ein	hjartesukk	over at vi enno ikkje h
n	n		2008	glitter i Jerusalem kjem eit lite	hjartesukk	frå Bernt Støylen : "De
n	n		2010	og testbildata . Det gjekk eit	sukk	gjennom testteamet .
n	n		2008	konto i banken . Eit lite	hjartesukk	i så måte , er at
m	m		1925	er-? Der læt ein	sukk	der , ein forunderleg u

4 Spørsmål eller kommentarar?

Takk for merksemda!

Referansar

Andreassen, Helene N.; Berez-Kroeker, Andrea; Collister, Lauren B.; Conzett, Philipp; Cox, Christopher; De Smedt, Koenraad; McDonnell, Bradley. 2019. Tromsø recommendations for citation of research data in linguistics. https://doi.org/10.15497/rda00040

Berez-Kroeker, Andrea L., Lauren Gawne, Barbara F. Kelly & Tyler Heston. 2017. A survey of current reproducibility practices in linguistics journals, 2003–2012. Henta 19. november frå https://sites.google.com/a/hawaii.edu/data-citation/survey.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. «Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field». *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Flick, Johanna. 2020a. Die Entwicklung des Definitartikels im Althochdeutschen. Eine kognitiv-linguistische Korpusuntersuchung (Empirically oriented theoretical morphology and syntax 6). Berlin: Language Science Press.

Flick, Johanna. 2020b. «Replication Data for: Die Entwicklung des Definitartikels im Althochdeutschen. Eine kognitiv-linguistische Korpusuntersuchung», https://doi.org/10.18710/HZKYL4, DataverseNO, V1.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017. «Putting practice into words: The state of data and methods transparency in grammatical descriptions». *Language Documentation & Conservation* 11. 157–189.

Hagen, Rune Blix. 2019. «Rettsforfulgte trollfolk i Finnmark, 1593-1692», https://doi.org/10.18710/OWP5IP, DataverseNO, V1.

Norsk Ordboks nynorskkorpus, http://no2014.uib.no/korpuset/.

Peng, Roger D. 2011. Reproducible Research in Computational Science. *Science*, 334, 6060. https://doi.org/10.1126/science.1213847.

Plesser, Hans E. 2018. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in neuroinformatics*, 11, 76. https://doi.org/10.3389/fninf.2017.00076.

Sollid, Hilde, Philipp Conzett & Åse Mette Johansen. 2014. «Gender and Noun Inflection: The Fate of 'Vulnerable' Categories in Northern Norwegian». *Studies in Language Variation* 16, 179–207. https://doi.org/10.1075/silv.16.09sol.

Thomason, Sarah. 1994. The editor's department. *Language* 70. 409–423.