UiT Noregs arktiske universitet

Gjenfinning og gjenbruk av forskingsdata – ei praktisk innføring i FAIR-prinsippa

KORG-dagane 2021, OsloMet/digitalt, 3.-4. juni 2021

Philipp Conzett
Universitetsbiblioteket
UiT Noregs arktiske universitet

ORCID: https://orcid.org/0000-0002-6754-7911
Twitter: @PhilippConzett

Takk for invitasjonen!

Oversikt over presentasjonen

Introduksjon

1. FAIR-prinsippa – kva og korfor?
2. Kva FAIR-prinsippa *ikkje* er
3. Gradar av FAIR

Korleis gjera data så FAIR som mogleg?

4. ... gjenfinnbare (Findable)
5. ... tilgjengelege (Accessible)
6. ... interoperable (Interoperable)
7. ... gjenbrukbare (Reusable)

Avslutning

8. Korleis kan du bidra med?

Spørsmål og diskusjon (ca. 5-10 min.)

Bolk 1:
Introduksjon

1. FAIR-prinsippa – kva og korfor?

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:

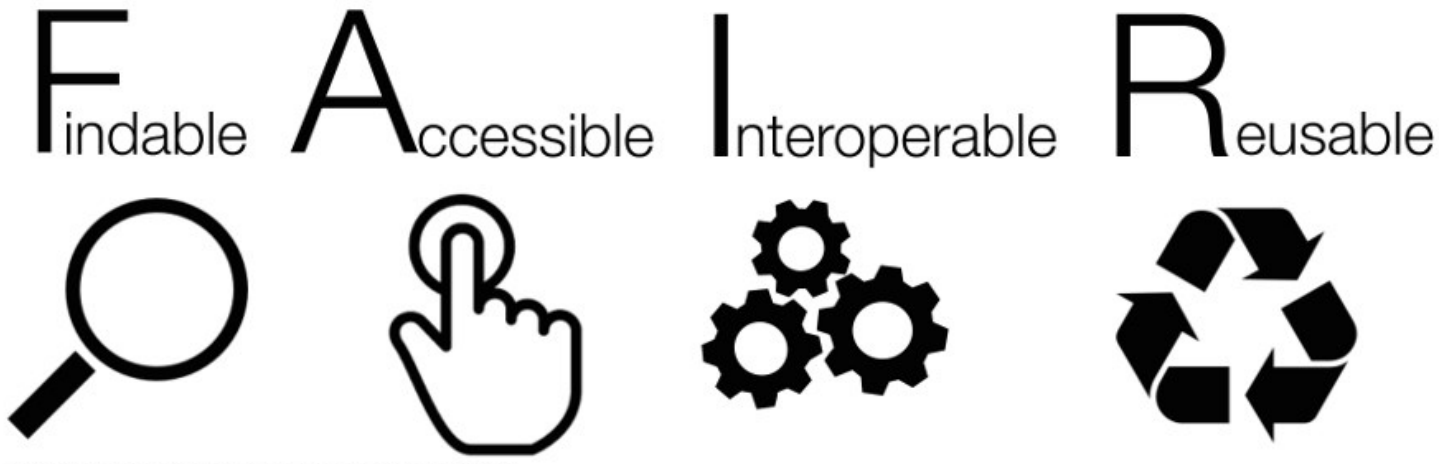


Image credit: Sungya Pundir, Wikimedia Commons CC BY-SA 4.0

1. FAIR-prinsippa – kva og korfor? (2)

Bakgrunn

- Introdusert i 2016 av Mark D. Wilkinson et al. i ein artikkel i *Scientific Data* som «The FAIR Guiding Principles for scientific data management and stewardship».
- Tek utgangspunkt i «an urgent need to improve the infrastructure supporting the reuse of scholarly data» (s. 1).

To overordna sider av FAIR:

- for menneske
- for maskinar (dvs. skal kunna handterast av maskinar; jf. «machine-actionability»


SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES

» Research data

» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

1. FAIR-prinsippa – kva og korfor? (3)

Korfor er det ønskjeleg å fremja gjenbruk av forskingsdata?

- Gjer det mogleg/lettare å **etterprøva** forskingsresultat og gjer dermed forskinga **meir transparent**.
- Kan føra til **nye forskingsresultat**.
- Reduserer dobbeltarbeid, og gjer dermed forskinga **meir effektiv**.
- Hjelper samfunnet med å takla **akutte ufordringar** som t.d. epidemiar.
- Aukar potensialet for **samarbeid** også på tvers av disiplinar.
- …

2. Kva FAIR-prinsippa *ikkje* er

FAIR-prinsippa har blitt interpreterte og brukte på ei rekkje ulike måtar. Nokre av forfattarane av FAIR-artikkelen frå 2016 kom derfor året etter med ei klargjering av kva FAIR er og kva det ikkje er (Mons et al. 2017).

Kva FAIR *er* …

- eit sett med **rettleiande prinsipp** som skal gje råd om korleis ein kan leggja til rette for at forskingsdata i aukande grad kan gjenbrukast (jf. s. 50)

2. Kva FAIR-prinsippa *ikkje* er (2)

Kva FAIR *ikkje* er …

- FAIR er **ikkje** ein **standard** eller ei **sjekkliste**. (jf. s. 51)

Derfor: Generelle forsikringar om at eit datasett, eit arkiv eller ei anna teneste «oppfyller FAIR-kriteria», «is FAIR-compliant», osb. bør takast med ei klype salt. Bruk heller formuleringar som «er i tråd med FAIR-prinsippa», «stør opp om FAIR-prinsippa», «is FAIR-aligned» eller liknande.

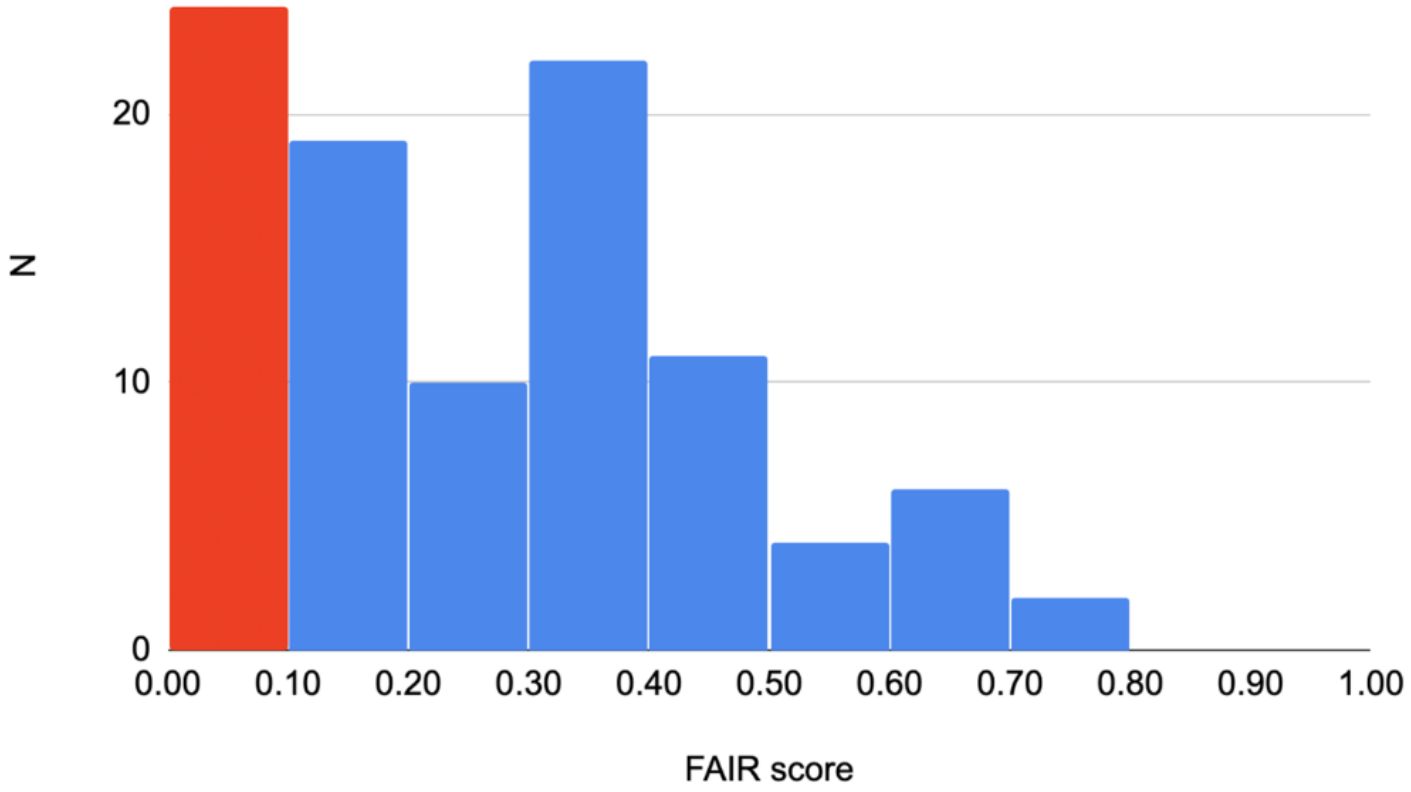
- FAIR data er **ikkje det same som opne data**. (jf. s. 51), men er i tråd med det generelle prinsippet «så ope som mogleg, men så lukka som nødvendig».

3. Gradar av FAIR

- Mons et al. (2017:50) skriv også: «[The FAIR principles] are a set of guiding principles that provide for a **continuum of increasing reusability**, via many different implementations.»
- FAIR er altså ikkje eit binært konsept, men forskingsdata kan vera **meir eller mindre FAIR**.
- Ulike aktørar held på å operasjonalisera FAIR-prinsippa og utvikla **konkrete kriterium** for å måla kor FAIR t.d. eit gjeve datasett er.
- Slike automatiserte FAIR-evalueringar er «berre» baserte på den **maskinhandterbare sida** av FAIR-prinsippa og undersøkjer i kva grad utvalde datasett i eit arkiv oppfyller dei operasjonaliserte FAIR-kriteria for ein maskin.

3. Gradar av FAIR (2)

EOSC-Nordic har FAIR-evaluert 74 forskingsdataarkiv i Norden og Baltikum. Her er ei grafisk oppsummering av resultatet (Jaunsen et al. 2020:23):



- Mesteparten av arkiva har ein FAIR-skår på mellom 10 og 50 %.
- Ingen av arkiva har ein FAIR-skår på meir enn 80 %.

Hugs! Dekkjer berre **den maskinhandterbare sida** av FAIR!

3. Gradar av FAIR

- Igjen: Grunn til å bli skeptisk når ein støyter på forsikringar om at eit datasett, eit arkiv eller ei anna teneste «oppfyller FAIR-kriteria», «is FAIR-compliant», osb.
- **Nokre fagområde**, som t.d. bioinformatikk, **ligg langt framme** når det gjeld FAIR. Her er det ofte store forskingsprosjekt som driv med dataintensiv forsking, og som har tilgang til avanserte fagspesifikke standardar og infrastrukturar som legg til rette for FAIR handtering av forskingsdata.
- Men så har vi heile den såkalla **lange halen av forskinga** som generer og handterer små eller mellomstore datasett, og der vi finn stor variasjon i datatypar og filformat. Slike fagområde manglar ofte etablerte fagspesifikke standardar for dokumentasjon, og infrastruktur for datahandtering.
- Det er først og fremst data av den siste typen eg skal fokusera på i resten av presentasjonen. Spørsmålet eg skal prøva å gje nokre svar på er: Korleis kan vi bidra til at **forskingsdata frå den lange halen av forskinga** blir **så FAIR som mogleg**?

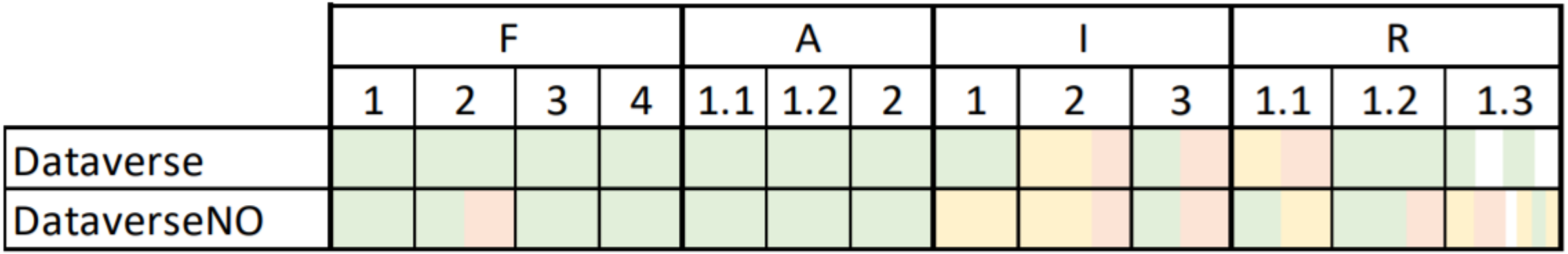
Bolk 2:
Korleis gjera data så FAIR som mogleg?

Eit kort og eit litt lengre svar …

- Det **korte svaret** er: Forskarane bør velja eit påliteleg arkiv allereie når dei lagar ein datahandteringsplan for prosjektet, og så følgja arkivet sine retningsliner (og andre beste-praksis-tilrådingar) for korleis data bør strukturerast og dokumenterast.
- Det litt **lengre svaret** blir ein litt meir detaljert, men kort gjennomgang av korleis dette kan sjå ut i praksis for dei enkelte elementa i FAIR (F, A, I, R).
- Gjennomgangen tek i stor grad **utgangspunkt i** ei støtteteneste som eg kjenner godt til og som eg jobbar med til dagleg, og det er **DataverseNO**, som er eit nasjonalt, generisk arkiv for forskingsdata frå forskarar frå norske forskingsinstitusjonar.
- Har undersøkt kor godt DataverseNO legg til rette for at data som er publiserte i arkivet, skal vera FAIR for både maskinar og menneske (jf. Conzett 2020). Resultatet kan oppsummerast slik:

Kor godt stør DataverseNO opp om FAIR?

Table 6: Summary of current FAIR implementation or support in Dataverse and DataverseNO



= (more or less) full implementation or support

= partial implementation or support

= (more or less) lacking implementation or support

= not applicable

Frå Conzett (2020:99)

La oss sjå på nokre utvalde døme på korleis DataverseNO stør opp om utvalde delar av FAIR-prinsippa.

Eit sentralt element i dette arbeidet er **retningslinene** som arkivet har, og frå ståstaden til forskaren er **arkiveringsguiden** som er viktig.

Arkiveringsguiden til DataverseNO

Før du arkiverer data i DataverseNO (inkl. dei ulike samlingane, t.d. UiT Open Research Data, TROLLing, osb.), må du sørgja for at dataa dine er i tråd med retningslinene nedanfor. DataverseNO aksepterer berre forskingsdata i digital form. God praksis for korleis ein skal førebu forskingsdata for arkivering kan oppsummerast slik:

- Bruk konsistente og forståelege filnamn (sjå bolk 1 nedanfor).
- Lagre dataa dine i eit føretrekt filformat (sjå bolk 2 nedanfor).
- Beskriv dataa dine i ei ReadMe-fil (sjå bolk 3 nedanfor).

Meir detaljerte retningsliner finn du nedanfor:

- **1 Filnamngjeving**
- **2 Føretrekte filformat**
- **3 Korleis beskriva dataa dine**

Korleis stør dette opp om FAIR?

>> La oss byrja med F-en i FAIR …

4. Korleis gjera forskingsdata så gjenfinnbare som mogleg?

- Publisert med persistent identifikator
- Gode metadata
- Indeksert

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

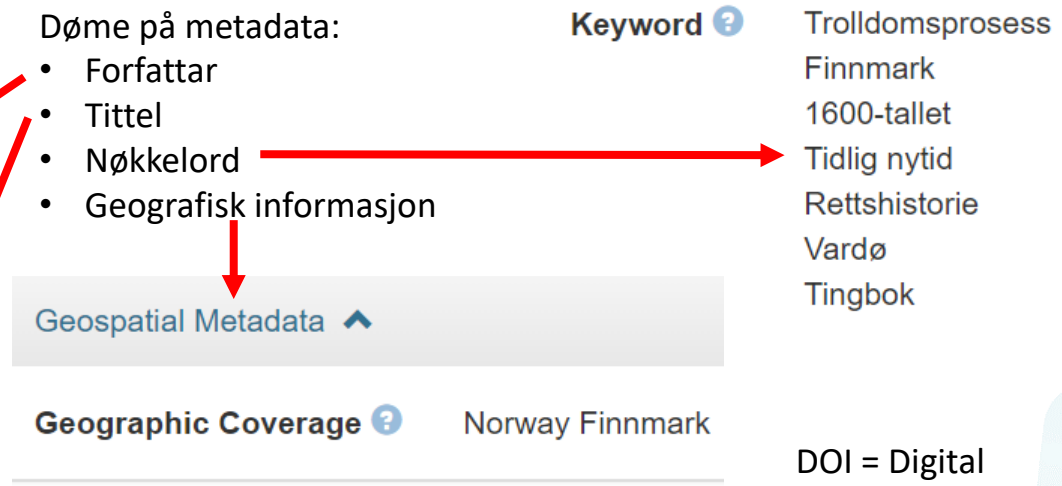
4. Korleis gjera forskingsdata så gjenfinnbare som mogleg?

- Publisert med persistent identifikator
- Gode metadata
- Indeksert

Metadata = beskriving av data

Døme på metadata:

- Forfattar
- Tittel
- 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

Metadata for gjennfinning i DataverseNO

Obligatoriske og tilrådde felt

Citation Metadata:

- ✓ **Title**
- ✓ **Author**, ORCID
- ✓ **Contact**
- ✓ **Description**
- ✓ **Keyword**
- ✓ Related Publication
- ✓ Language
- ✓ Contributor

Citation Metadata:

- ✓ Grant Information
- ✓ Time Period Covered
- ✓ Date of Collection
- ✓ Kind of Data
- ✓ Related Material
- ✓ Related Dataset
- ✓ Data Sources

Geospatial Metadata:

- ✓ Geographic Coverage
- ✓ Geographic Bounding Box

+ fleire valfrie felt

...

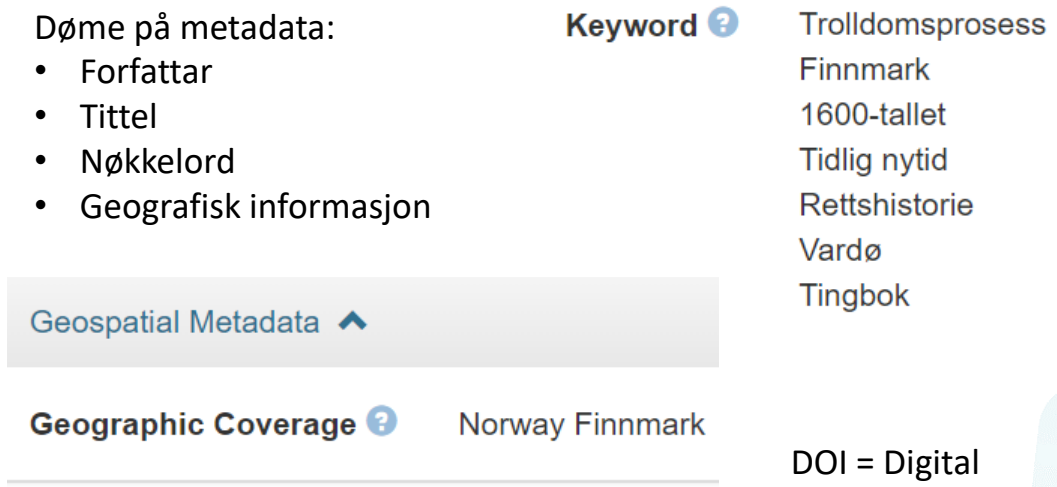
Særleg relevant for gjenfinning

- Publisert med persistent identifikator
- Gode metadata
- Indeksert

Metadata = beskriving av data

Døme på metadata:

- Forfattar
- Tittel
- Nøkkelord
- Geografisk informasjon



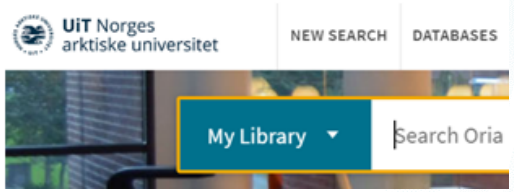
Hagen, Rune Blix, 2019, "Rettsforfulgte trollfolk i Finnmark, 1593-1692", https://doi.org/10.18710/OWP5IP, DataverseNO, V1

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

DataCite
FIND, ACCESS, AND REUSE DATA

Google Dataset Search Beta

BASE
Bielefeld Academic Search Engine



5. Korleis gjera forskingsdata så tilgjengelege så mogleg?

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 gjennom ein veldefinert og open protokoll (t.d. https).

- Veldefinert og open protokoll
- Tilstrekkeleg autentisering

- Ein bør også velja eit arkiv som gjer dataa tilgjengelege i tråd med innhaldet. Døme:

  Visse typar data kan ikkje gjerast ope tilgjengelege (t.d. ikkje-anonymiserte persondata), men dei kan kanskje likevel delast i eit arkiv der dei som ønskjer å lasta ned data, må registrera seg og logga inn. Då treng ein tilstrekkeleg autentisering.

6. Korleis gjera forskingsdata så interoperable som mogleg?

- Opne metadata-format
- Felles standardar
- Konsistente vokabular

- 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 innanfor samfunnsfag og helsefag.
- Bruk **konsistente metadatavokabular**, t.d. DDI-vokabularet for aggregeringsmetode (Aggregation Method); utdrag:
- Interoperabilitet mogleg-gjer **søk og gjenbruk på tvers av datasett og arkiv**.

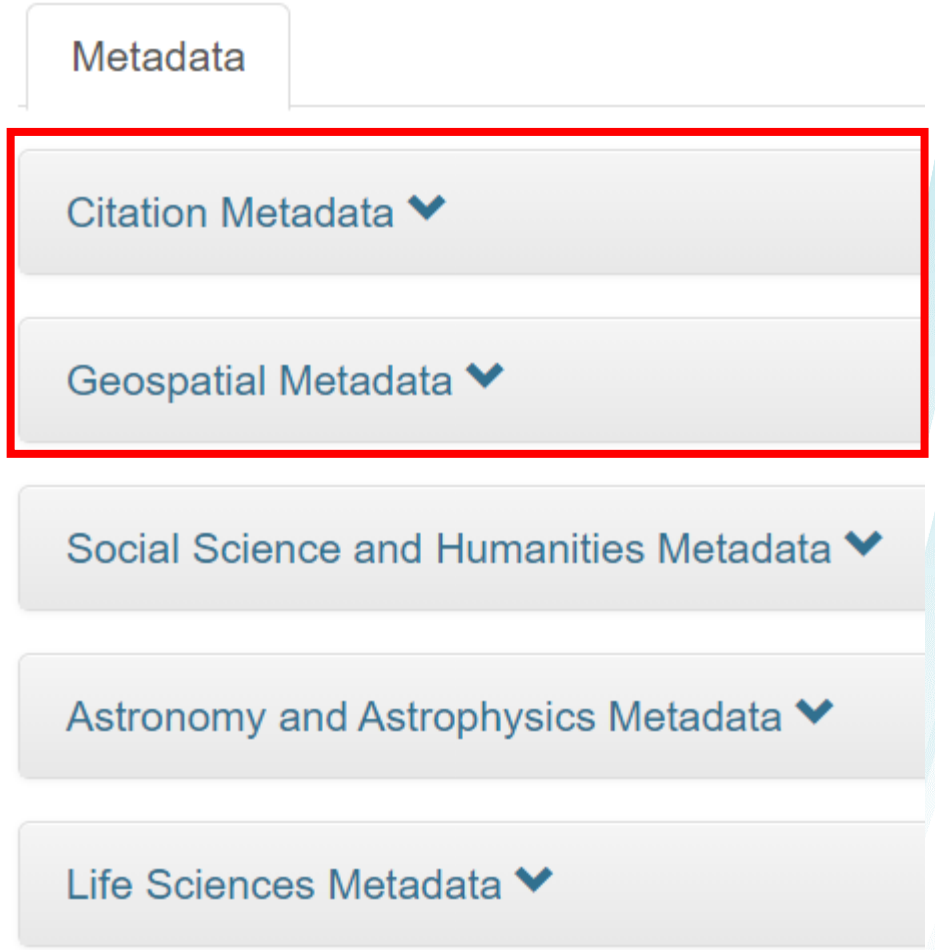
Value of the Code	Descriptive Term of the Code	Definition of the Code
Maximum	Maximum	The highest value attained or recorded.
Minimum	Minimum	The lowest value attained or recorded.

Støtte for interoperabilitet i DataverseNO

- I-en i FAIR er kanskje det svakaste punktet i DataverseNO (og mange andre, særleg generiske arkiv).
- Men noka støtte er på plass både for metadatastandardar og kontrollerte vokabular.

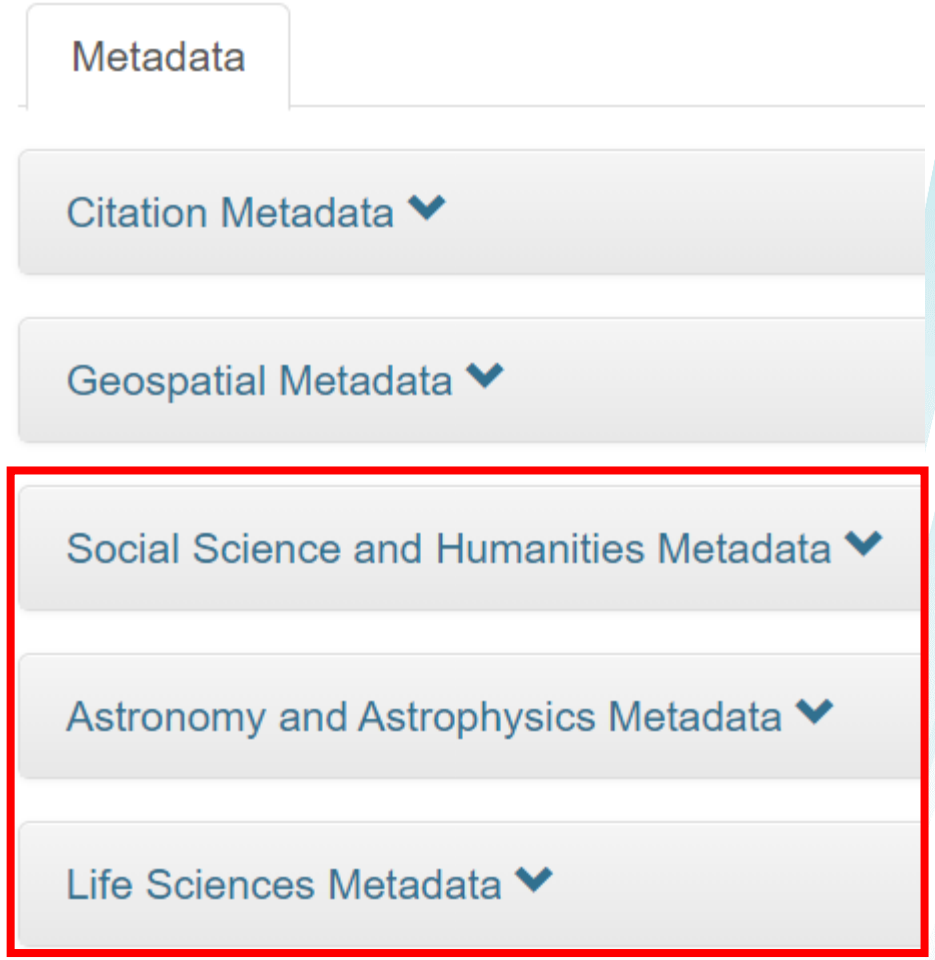
Metadatastandardar i DataverseNO

- Bruker Dublin Core og delar av DDI og delar av ISO for generelle metadata («Citation Metadata») og geografiske metadata.



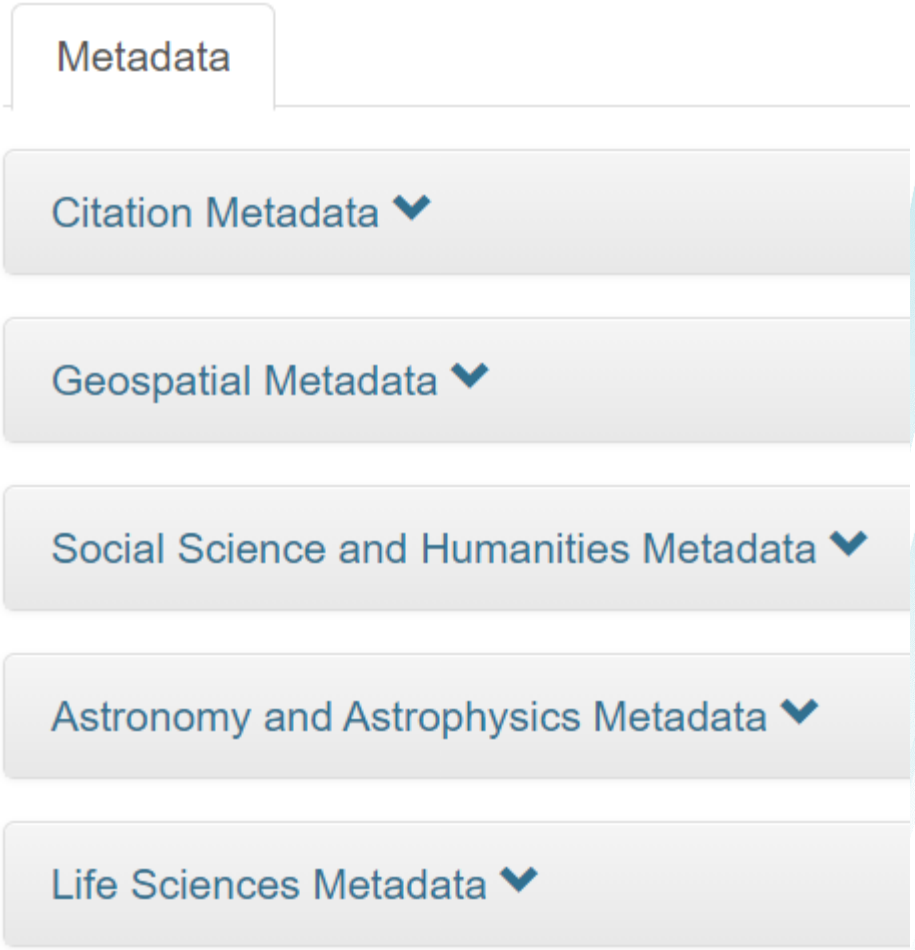
Metadatastandardar i DataverseNO

- Bruker Dublin Core og delar av DDI og delar av ISO for generelle metadata («Citation Metadata») og geografiske metadata.
- Har tre skjema for meir fagspesifikke metadata for 1) samfunnsvitskap og humaniora; 2) astronomi og astrofysikk; og 3) biovitskap.



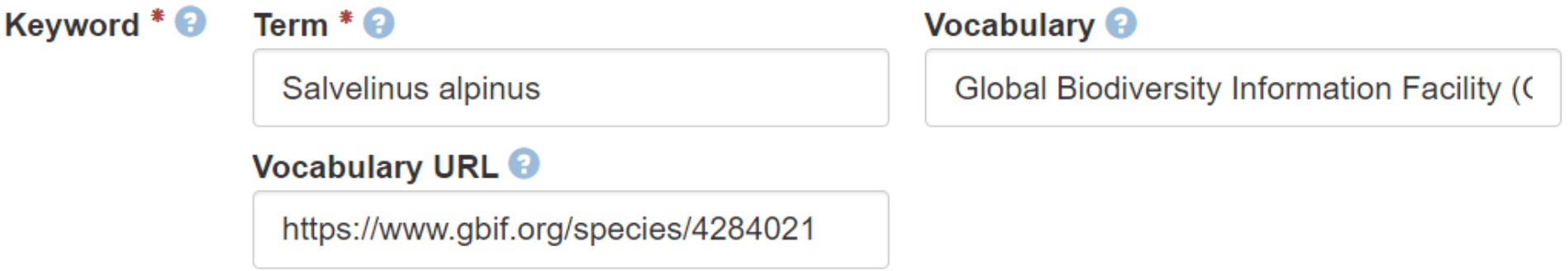
Metadatastandardar i DataverseNO

- Bruker Dublin Core og delar av DDI og delar av ISO for generelle metadata («Citation Metadata») og geografiske metadata.
- Har tre skjema for meir fagspesifikke metadata for 1) samfunnsvitskap og humaniora; 2) astronomi og astrofysikk; og 3) biovitskap.
- Skjema for fleire fagspesifikke metadatastandardar er under utvikling, t.d. CESSDA Metadata Model (samfunnsvitskap), Darwin Core (biofag), CMDI (språkfag).



Kontrollerte vokabular i DataverseNO

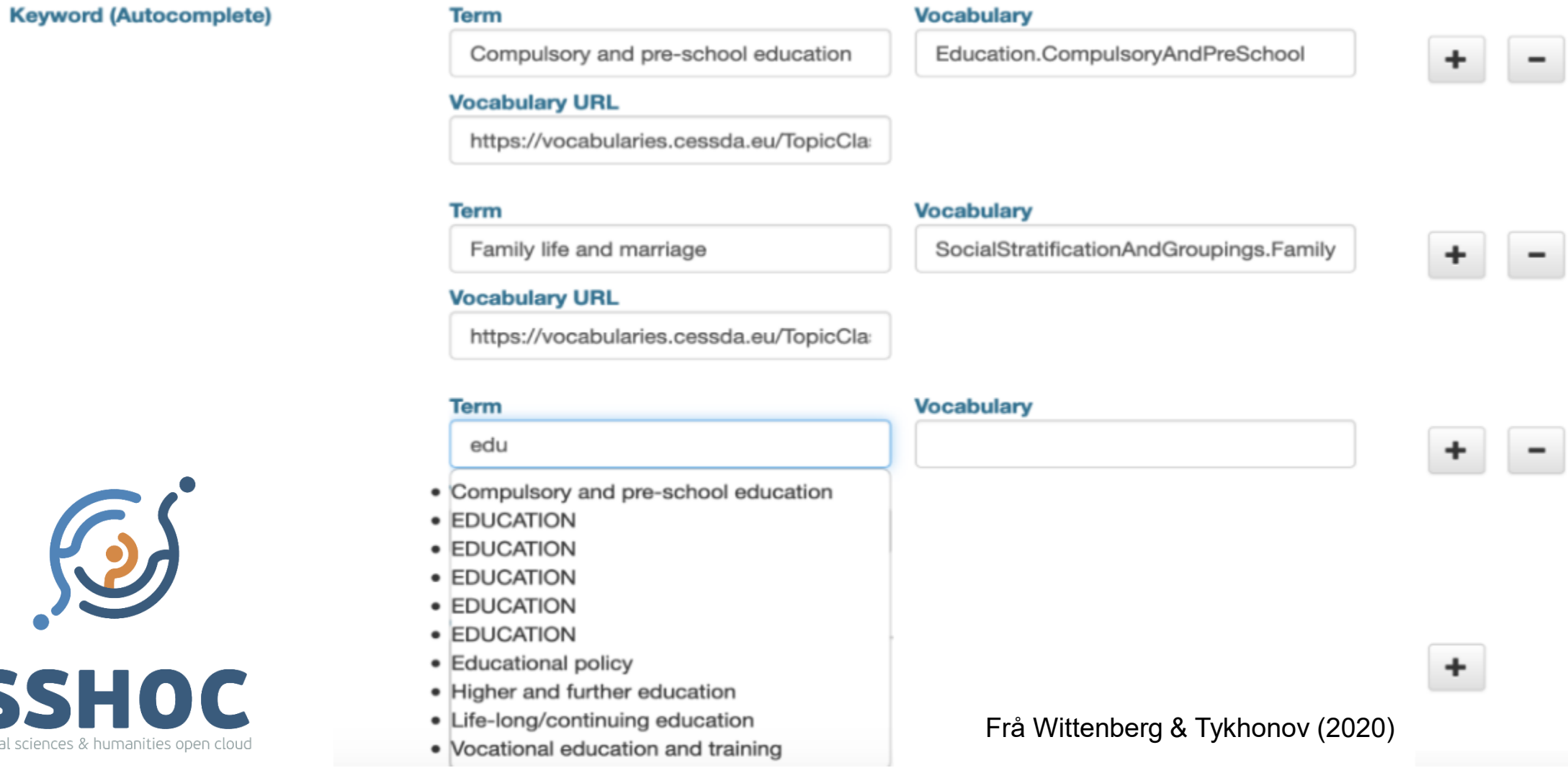
- Informasjon om kontrollerte vokabular kan leggjast til som fritekst på nøkkelord:



- … er ikkje maskinhandterbart, men …

- … betre/full støtte for kontrollerte vokabular er under utvikling som del av eit EU-prosjekt som heiter SSHOC (Social Sciences and Humanities Open Cloud). Målet er at ein skal kunna velja eit kontrollert vokabular frå ein nedtrekksmeny, og deretter velja ein eller fleire verdiar frå vokabularet:

Kontrollerte vokabular i DataverseNO (2)



SSHOC
social sciences & humanities open cloud

Frå Wittenberg & Tykhonov (2020)

7. Korleis gjera forskingsdata så gjenbrukbare som mogleg?

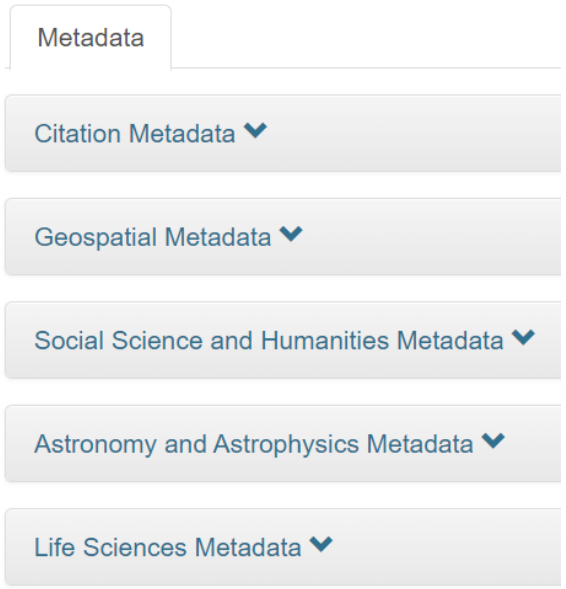
- Dokumenter data, slik at dei er forståelege og kan gjenbrukast av fagfellar.
- Arkiver data i føretrekte/arkivverdige filformat slik at filene kan opnast og lesast på lang sikt. Døme: For tabelldata, bruk rein tekst (.txt) i tillegg til eller i staden for Excel (.xlsx).
- Definer ein klar brukslisens 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.

Dokumentasjon i DataverseNO

3 Korleis beskriva dataa dine

For at andre forskarar skal kunna forstå og gjenbruka dataa dine er det viktig at du beskriv dei på ein konsistent og forståeleg måte før dei blir publiserte. Det er to plassar i DataverseNO der du skal leggja inn slik dokumentasjon, i **metadatafelta** og i ei separat **ReadMe-fil** som skal lastast opp saman med datafilene:

- **Metadata**
- **ReadMe-fil**



ReadMe-fil

Forklaring på korfor det er viktig med god dokumentasjon >> gjenbruk!

ReadMe-fil

Ei **ReadMe-fil** er ei meir detaljert rettleiing på datasettet ditt som gjer det mogleg for andre forskarar å tolka, forstå og gjenbruka dataa dine. ReadMe-fila dokumenterer korleis datasettet er oppretta, kor fullstendig det er, og kva slags begrensningar det har. ReadMe-fila må minimum innehalda dette:

- Tittel på datasettet, DOI, kontaktinformasjon
- Metode
- Data- og filoversikt
- Filspesifikk informasjon
- Vilkår for gjenbruk

Minstekrava til ei ReadMe-fil

Lenkje til ReadMe-filmal

Bruk gjerne denne **generelle malen** som utgangspunkt for ReadMe-fila.

ReadMe-filmal

… med rettleiing/hjelp til forskaren

GENERAL INFORMATION

METHODOLOGICAL INFORMATION

<Note! It may generally be considered appropriate to have **overlap** in the methods section of a research data README file **with** citation of the **original article**. See Committee on Publication Ethics (COPE) guidance on text recycling: https://...>

DATA & FILE OVERVIEW

DATA-SPECIFIC INFORMATION FOR: [FILENAME]

SHARING/ACCESS INFORMATION

ReadMe-fil (2)

ReadMe-fil

Ei **ReadMe-fil** er ei meir detaljert rettleiing på datasettet ditt som gjer det mogleg for andre forskarar å tolka, forstå og gjenbruka dataa dine. ReadMe-fila dokumenterer korleis datasettet er oppretta, kor fullstendig det er, og kva slags begrensningar det har. ReadMe-fila må minimum innehalda dette:

- Tittel på datasettet, DOI, kontaktinformasjon
- Metode
- Data- og filoversikt
- Filspesifikk informasjon
- Vilkår for gjenbruk

Lenkjer til autentiske ReadMe-filer

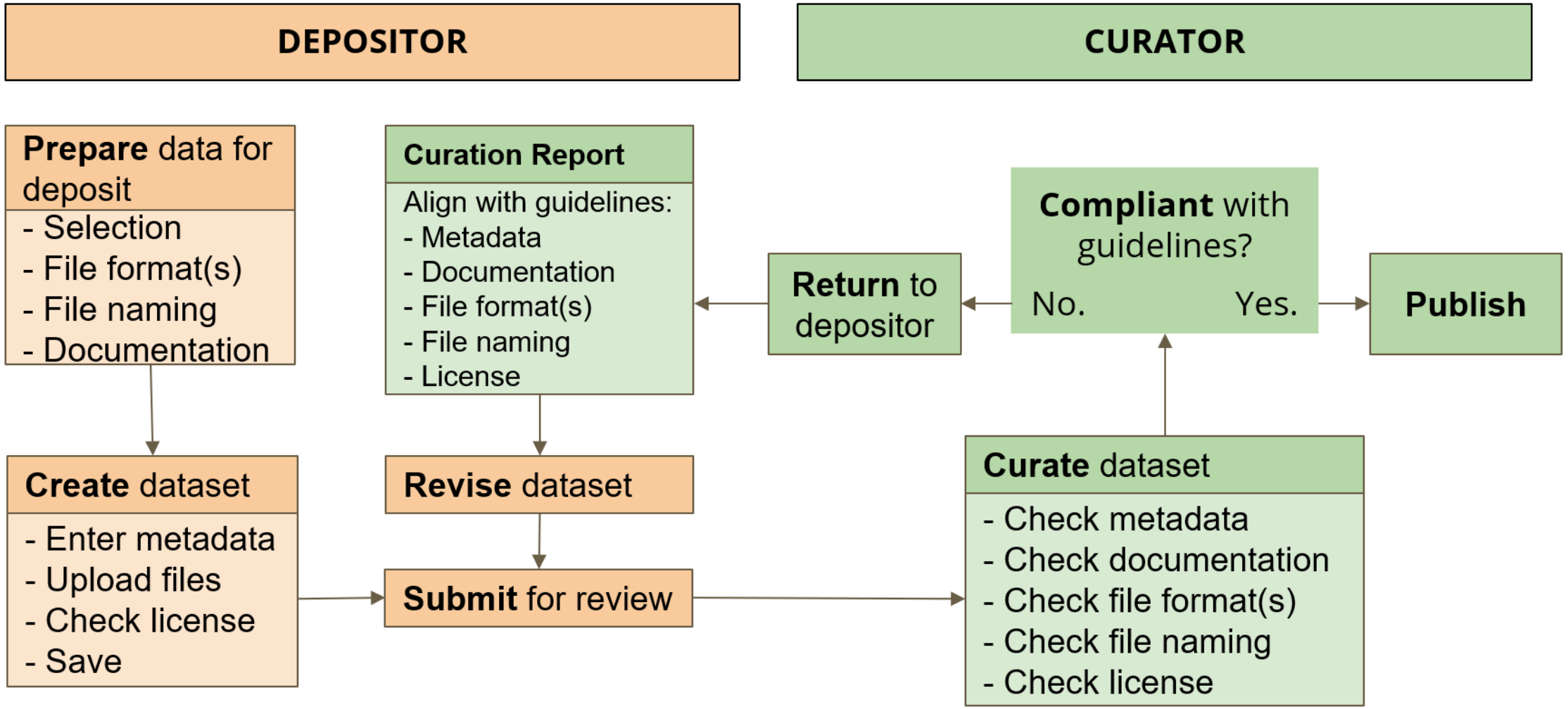
Bruk gjerne denne **generelle malen** som utgangspunkt for ReadMe-fila.

Her er nokre døme på ReadMe-filer: døme 1 (samfunnsfag); døme 2 (naturvitskap).

Kvalitetskontroll i DataverseNO

- Så langt har vi snakka om korleis vi kan guida forskarar til å gjera dataa sine så FAIR som mogleg.
- Men gode guidar er ofte ikkje nok.
- Derfor går alle datasetta gjennom **kuratering** før publisering.
- Inga tid til å gå gjennom dette her og no, men prosessen kan oppsummertast slik:

Deponering, kuratering og publisering



Bolk 3:
Avslutning

8. Korleis kan du bidra til at forskingsdata på institusjonen din blir så FAIR som mogleg?

Som bibliotekar:

- Sørgja for at forskarar finn relevant informasjon om forskingsdatahandtering på nettsidene til biblioteket. NB! Gjenbruk! T.d. Open Science Toolbox (http://openscience.prototyp.io/).

Som bibliotekdirektør:

- Byggja opp kompetanse og støttetenester.
- Byggja på eksisterande ressursar, t.d. OA-kompetanse, fagreferentar.

Som dekan/instituttleiar/…:

- Gjera forskarane merksame på støttetenester som biblioteket og andre tilbyr.
- Gjera det meir attraktivt for forskarar å bruka tid på å gjera dataa sine så FAIR som mogleg (jf. insentiv til forskingstermin m.m.).

Som rektor/prorektor/ forskings-direktør/…:

- Få på plass ein policy for forskingsdatahandtering.
- Sørgja for at forskarar har tilgang til nødvendig infrastruktur.

Som forskar:

- Følgja tilrådingar innanfor faget ditt.
- Få biblioteket med på laget.

Meld dokker inn i den norske RDA-noden: https://rd-alliance.org/groups/rda-norway**!**

Takk for merksemda!

Bolk 4:
Spørsmål og diskusjon

Referansar

Conzett, Philipp. 2020. «DataverseNO: A National, Generic Repository and Its Contribution to the Increased FAIRness of Data from the Long Tail of Research». *Ravnetrykk*, nr. 39: 74–113. https://doi.org/10.7557/15.5514.

Jaunsen, Andreas Ortmann, Mari Kleemola, Tuomas J. Alaterä, Heikki Lehvaslaiho, Adil Hasan, Josefine Nordling, og Pauli Assinen. 2020. «D4.1 An Assessment of FAIR-Uptake among Regional Digital Repositories», august. https://doi.org/10.5281/zenodo.4045402.

Mons, Barend, Cameron Neylon, Jan Velterop, Michel Dumontier, Luiz Olavo Bonino da Silva Santos, og Mark D. Wilkinson. 2017. «Cloudy, Increasingly FAIR; Revisiting the FAIR Data Guiding Principles for the European Open Science Cloud». *Information Services & Use* 37 (1): 49–56. https://doi.org/10.3233/ISU-170824.

Om DataverseNO. https://site.uit.no/dataverseno/nn/om/.

SangyaPundir. 2016. File:FAIR data principles.jpg. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, mfl. 2016. «The FAIR Guiding Principles for Scientific Data Management and Stewardship». *Scientific Data* 3: 160018. https://doi.org/10.1038/sdata.2016.18.

Wittenberg, Marion & Vyacheslav Tykhonov. 2020. Dataverse in the European Open Science Cloud. *Septentrio Conference Series*, 2 (2020). https://doi.org/10.7557/5.5421.