OSLOMET

KANDIDAT

303

PRØVE

MBIB5900 1 Masteroppgave

Emnekode	MBIB5900
Vurderingsform	Individuell eksamen uten tilsyn
Starttid	11.06.2020 10:00
Sluttid	16.06.2020 10:00
Sensurfrist	28.07.2020 23:59
PDF opprettet	03.11.2020 14:19
Opprettet av	Ragnhild Hove

Forfatter

Navn på forfatter (etternavn, fornavn):

Skriv tekst her

Holtet, Vigdis

Besvart.

Veileder

Navn på veileder (etternavn, fornavn). Har du hatt biveileder, oppgi navn på både hovedveileder og biveileder:

Skriv tekst her

Pharo, Nils

Besvart.

Tittel på oppgaven

Tittel på oppgaven (nøyaktig slik den er skrevet i oppgaven, med undertittel):

Skriv tekst her

Kvaliteten på norske metadata for forskningsdata

Besvart.

Innleveringsår

Årstall innlevert:

Velg ett alternativ

- 2019
- 2020
- 2021
- 2022
- 2023
- 2024
- 2025
- 2026
- 2027
- 2028

Besvart.

Fakultet

Ved hvilket fakultet ved OsloMet leveres masteroppgaven?

Velg ett alternativ

- Fakultet for helsevitenskap
- Fakultet for lærerutdanning og internasjonale studier
- Fakultet for samfunnsvitenskap
- Fakultet for teknologi, kunst og design

Besvart.

Abstract/sammendrag

Abstract/sammendrag (kopi av abstract i oppgaven der dette finnes) ca. 300 ord:

Skriv tekst her

Målet med denne oppgaven er å undersøke hva som kjennetegner kvaliteten på norske metadata for forskningsdata. Både nasjonalt og internasjonalt er det forventet at forskere skal gjøre sine data digitalt tilgjengelige for replisering og gjenbruk. For at dette skal kunne gjennomføres, er det viktig at dataene er lagret ved hjelp av metadata som er av høy kvalitet. FAIR-prinsippene, som er et akronym som uttrykker at forskningsdata og metadata skal være Findable, Accessible, Interoperable og Reusable, er velegnet for å måle denne kvaliteten. FAIR-prinsippene, i kombinasjon med sentral litteratur om metadata-kvalitet, har derfor blitt benyttet til å utviklet et analyseverktøy. Et representativt utvalg av metadata er blitt høstet fra et norsk generisk forskningsdata-arkiv og analysert ved hjelp av analyseverktøyet. Resultatene stemmer godt overens med det som finnes av tidligere forskning på området. Metadataene har flere viktige egenskaper som legger til rette for gjenfinning og gjenbruk, og som dermed styrker kvaliteten. De har imidlertid også vesentlige svakheter som alle har med manglende standardisering å gjøre, og som særlig hindrer mulighetene for interoperabilitet. Det anbefales derfor å forske videre på dette området, blant annet ved å utvikle kontrollerte vokabularer for de ulike disipliner.

Besvart.

Emneord

Emneord/stikkord relatert til emnet masteroppgaven omhandler (Hvert emneord skrives med stor forbokstav ellers små bokstaver):

Skriv tekst her

Forskningsdata
FAIR metadata
Metadata-kvalitet
Norge

Besvart.

Semesteradresse

Korrekt kontaktinformasjon er en forutsetning for at du skal få tilsendt sluttdokumentasjon.

Jeg har sjekket at min semesteradresse og telefonnummer i Studentweb er korrekt:

Velg ett alternativ

- (•) Ja
- ( ) Nei

Besvart.

1

NSD/REK

Dersom du behandler personopplysninger i masteroppgaven din, skal dette være meldt til NSD (Norsk senter for forskningsdata). I tillegg kan det være nødvendig med godkjenning fra REK (Regionale komitéer for medisinsk og helsefaglig forskningsetikk).

Velg ett alternativ

- Jeg har ikke meldt prosjektet til NSD/REK. ✔
- Jeg har meldt prosjektet til NSD og fått det godkjent. ✖
- Jeg har meldt prosjektet til NSD og REK og fått det godkjent.
- Jeg har meldt prosjektet til NSD, men ikke fått svar eller godkjenning.
- Jeg har meldt prosjektet til NSD og REK, men ikke fått svar eller godkjenning.

Feil. 0 av 0 poeng.

2

Klausulering

Dersom oppgaven er klausulert, må klausuleringsskjema leveres inn sammen med oppgaven. Oppgaver som klausuleres etter lovbestemt taushetsplikt (§ 13 i Forvaltningsloven) skal ikke publiseres i ODA. Klausulering av andre grunner kan gjøres i maksimalt 5 år.

Klikk for å lese mer om personopplysninger og personvern

Lenke til klausuleringsskjemaene

Velg ett alternativ

- Nei ✔
- Ja, grunnet lovpålagt taushetsplikt
- Ja, av andre grunner i 1 år (embargotid 1 år)
- Ja, av andre grunner i 2 år (embargotid 2 år)
- Ja, av andre grunner i 3 år (embargotid 3 år)
- Ja, av andre grunner i 4 år (embargotid 4 år)
- Ja, av andre grunner i 5 år (embargotid 5 år)

Riktig. 0 av 0 poeng.

3

Lisensavtale

Ved å godta denne avtalen gir du som forfatter Universitetsbiblioteket (UB), på vegne av OsloMet - storbyuniversitetet (tidligere Høgskolen i Oslo og Akershus), rett til vederlagsfritt å gjøre det innleverte dokumentet tilgjengelig i et institusjonelt arkiv.

- **Denne avtalen er ikke til hinder for at forfatteren kan publisere dokumentet i papirform eller elektronisk annetsteds i en identisk eller endret versjon.**
- **Forfatteren skal avlevere dokumentet i et maskinleselig format spesifisert av OsloMet.**
- **Forfatteren garanterer at han eller hun har opphavsrett til arbeidet, eller tillatelse fra rettighetshaverne til å publisere teksten på internett.**
- **OsloMet har ikke adgang til kommersiell utnytting av dokumentet.**
- **Arbeidet kan brukes, det vil si vises til, lenkes til, siteres fra, skrives ut og lastes ned innenfor de rammer som lov om opphavsrett til åndsverk av 15. Juni 2018 (åndsverkloven) med endringer angir. (https://lovdata.no/dokument/NL/lov/2018-06-15-40)**

Forfatteren må skriftlig søke UB om å si opp avtalen.

Det gjøres oppmerksom på at masteroppgaver som utgangspunkt er offentlige. Dersom oppgaven ikke er klausulert eller embargotiden har utløpt, vil det kunne gis innsyn i masteroppgaver som ligger i universitetets ordinære arkiv (pt. Public 360) selv om oppgaven ikke er publisert i ODA.

Velg ett alternativ

- (•) JA, jeg godtar avtalen og vil at min masteroppgave skal være tilgjengelig for omverden (er oppgaven klausulert en viss periode av andre grunner, blir den tilgjengelig når embargotiden er utløpt) ✓
- ( ) NEI, jeg vil ikke at min masteroppgave skal være tilgjengelig for omverden (er oppgaven klausulert pga. taushetsplikt må du velge dette svaralternativet)

Riktig. 0 av 0 poeng.

4

Innlevering klausuleringsskjema

Innlevering klausuleringsskjema

Hvis oppgaven ikke er klausulert, klikk deg videre til innlevering av masteroppgaven.

Last opp filen her. Maks én fil.

Følgende filtyper er tillatt: **.pdf** Maksimal filstørrelse er **5 GB**.

Velg fil for opplasting

Ubesvart.

5 # Innlevering av masteroppgave

Innlevering masteroppgave

Din fil ble lastet opp og lagret i besvarelsen din.

Last ned | Fjern | Erstatt

Filnavn:	VigdisHoltet.Master2020.pdf
Filtype:	application/pdf
Filstørrelse:	1.82 MB
Opplastingstidspunkt:	11.06.2020 11:43
Status:	**Lagret**

Besvart.

Vigdis Holtet

Kvaliteten på norske metadata for forskningsdata

Masteroppgave 2020

Master i bibliotek- og informasjonsvitenskap

OsloMet — Storbyuniversitetet, Institutt for arkiv-, bibliotek- og informasjonsfag

OSLO METROPOLITAN UNIVERSITY
STORBYUNIVERSITETET

Sammendrag

Målet med denne oppgaven er å undersøke hva som kjennetegner kvaliteten på norske metadata for forskningsdata. Både nasjonalt og internasjonalt er det forventet at forskere skal gjøre sine data digitalt tilgjengelige for replisering og gjenbruk. For at dette skal kunne gjennomføres, er det viktig at dataene er lagret ved hjelp av metadata som er av høy kvalitet. FAIR-prinsippene, som er et akronym som uttrykker at forskningsdata og metadata skal være Findable, Accessible, Interoperable og Reusable, er velegnet for å måle denne kvaliteten. FAIR-prinsippene, i kombinasjon med sentral litteratur om metadata-kvalitet, har derfor blitt benyttet til å utviklet et analyseverktøy. Et representativt utvalg av metadata er blitt høstet fra et norsk generisk forskningsdata-arkiv og analysert ved hjelp av analyseverktøyet. Resultatene stemmer godt overens med det som finnes av tidligere forskning på området. Metadataene har flere viktige egenskaper som legger til rette for gjenfinning og gjenbruk, og som dermed styrker kvaliteten. De har imidlertid også vesentlige svakheter som alle har med manglende standardisering å gjøre, og som særlig hindrer mulighetene for interoperabilitet. Det anbefales derfor å forske videre på dette området, blant annet ved å utvikle kontrollerte vokabularer for de ulike disipliner.

Summary

The aim of this thesis is to explore the characteristics of the quality of Norwegian metadata for research data. Both nationally and internationally it is expected that researchers make their data digitally available for replication and re-use. In order for this to be accomplished, it is important that the research data are stored accompanied by high quality metadata. The FAIR-principles, which are an acronym denoting that data and metadata should be Findable, Accessible, Interoperable and Reusable, are a suitable measurement for this quality. The FAIR-principles, combined with pertinent literature about metadata quality, have therefore been used to develop a tool for analysis. A representative selection of metadata has been harvested from a Norwegian generic research data archive and analysed. The results conform to previous research on the quality of metadata for research data. The metadata have several attributes that facilitate findability and re-use, hence improving the quality. They have, however, also significant weaknesses that are all connected to the lack of standardization, and that primarily prevent interoperability. Further research in this area is therefore reccommended, one possible focus being the development of controlled vocabularies within different disciplines.

OsloMet — Storbyuniversitetet, Institutt for arkiv-, bibliotek- og informasjonsfag

Oslo 2020

Innhold

Figurer

Tabeller

1 Innledning

Økende grad av digitalisering har gjort det mulig å lagre, dele og gjenbruke forskningsdata på en helt annen måte enn tidligere. Open Access/åpen tilgang[1] innebærer at vitenskapelige resultater i form av artikler og rapporter skal være fritt tilgjengelige, men det innebærer også fri tilgang til forskningsdataene som ligger til grunn for publikasjonene. Slik kan forskningsdataene bli ressurser som kommer til nytte for andre forskere og studenter, samtidig som forskningen kan etterprøves og kvalitetssikres. Dessuten unngår offentlige institusjoner å finansiere den samme datainnsamlingen flere ganger. Tilgang på forskningsdata gir også mulighet for økt innovasjon, ved at andre aktører enn forskere får tilgang til dataene. I tillegg kan innsyn i forskningsdata øke tilliten til forskere og forskningen (Kunnskapsdepartementet, 2017, s. 7; RDA-Codata Legal Interoperability Interest Group, 2016).

Regjeringen slo i desember 2017 fast at "[f]orskning som skjer ved bruk av offentlige midler, skal være til det beste for alle." Videre introduserte de tre grunnprinsipper for deling av forskningsdata: (1) Forskningsdata skal være så åpne som mulig, og så lukkede som nødvendig ; (2) Forskningsdata bør håndteres og tilrettelegges slik at verdiene i dataene kan utnyttes best mulig ; (3) Beslutninger om arkivering og tilrettelegging av forskningsdata må tas i forskerfellesskapene (Kunnskapsdepartementet, 2017, s. 23-26)

Forskningsdata kan defineres som

> factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated. (OECD, 2007, s. 13)

Her impliseres hvilke typer forskningsdata som finnes, at de er førstehånds kilder for forskningen, og at de er allment akseptert i de vitenskapelige miljøene som nødvendige for å validere forskningsresultatene. Dessuten slås det fast at datasettet er representert på en systematisk måte og dekker deler av emnet det forskes på.

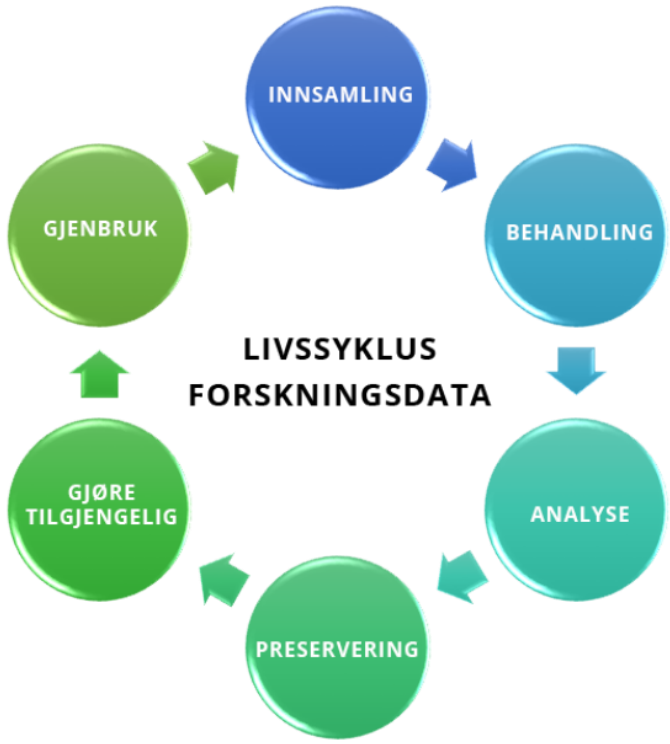
[1] https://www.openaccess.no/

To viktige forutsetninger kreves for at forskningsdata skal kunne deles og gjenbrukes: For det første er det nødvendig med en infrastruktur for deling, det vil si forskningsstøtte-tjenester som kan bistå med kuratering av dataene, samt forskningsdata-arkiver hvor data kan lagres og være tilgjengelige over tid. For det andre må data og metadata være standardiserte, slik at de kan prosesseres og gjenfinnes (Day, 2010, s. 15). Internasjonalt finnes det mange arkiver for forskningsdata, såkalte *research data repositories*. Nettstedet Registry of Research Data Repositories[2] gir en samlet søkbar oversikt over de fleste av disse arkivene. Eksempler på arkiver man kan finne der, kan være nasjonale generiske arkiver eller arkiver innen en bestemt disiplin. I Norge har vi fem generiske arkiver for forskningsdata som de ulike forskningsinstitusjonene kan benytte for å lagre sine data. Disse vil bli nærmere introdusert i oppgavens metode-kapittel.

Fra innsamling til gjenbruk går forskningsdata gjennom et kretsløp som er illustrert i Figur 1 nedenfor. Prosessen begynner med forskerens innsamling av data, innledet av utformingen av en datahåndteringsplan (DMP)[3]. I en slik plan må forskeren dokumentere hva slags data som skal samles inn, hvordan innsamlingen skal foregå, hvilke metadata som skal brukes, hvordan etiske og juridiske problemstillinger skal håndteres, hvilke data som blir lagret under prosjektet og hvilke som lagres ved prosjektets slutt, og hvem som skal ha tilgang til dataene under og etter prosjektet (NSD, u.å.).

Etter at dataene er samlet inn, behandles/prosesseres de, for eksempel ved transkribering, rensing, anonymisering, lagring m.m. Neste steg er analyse av dataene, det vil si at selve forskningsarbeidet utføres. Så preserveres dataene, noe som innebærer lagring, back-up, transformering til egnet filformat, og utforming av metadata. Deretter gjøres dataene tilgjengelige i et søkbart arkiv og med de nødvendige tilganger for gjenbruk. Dataene kan så gjenbrukes, slik at videre forskning eller ny forskning kan gjøres på dem, forskning kan etterprøves og de kan brukes i undervisningsøyemed. Gjennom gjenbruk kan nye data skapes, og kretsløpet går videre (University of Vienna, u.å.).

[2] https://www.re3data.org/
[3] DMP – Data Management Plan



Figur 1: Livssyklus for forskningsdata[4]

Det finnes flere aspekter ved temaet tilgjengeliggjøring og gjenbruk av forskningsdata som kan være gjenstand for forskning. Noen har undersøkt forskningsdata-arkiver innen ulike disipliner for å kartlegge rutiner for deponering, bruk av ulike metadata-skjemaer m.m. Dette har med infrastrukturen for forskningsdataene å gjøre (Austin et al., 2016; Kim et al., 2019). Andre har analysert kvaliteten på metadataene som forskningsdataene lagres ved hjelp av (Balatsoukas et al., 2018; Rousidis et al., 2014; Rousidis et al., 2015).

En viktig forutsetning for at forskningsdata skal kunne gjenfinnes og gjenbrukes, er at de er lagret ved hjelp av metadata som legger til rette for dette, og fokus for denne oppgaven er kvaliteten på disse metadataene. Som Austin et al. (2016) slår fast, er metadata "the backbone of any dataset, and ongoing quality control of metadata is as important as the data" (s. 27). Med andre ord er det ikke tilstrekkelig å etterstrebe høy kvalitet på selve forskningsdataene. Det er like viktig at metadataene er av høy kvalitet. De to henger uløselig sammen, og forskningsdata kan verken lagres, gjenfinnes eller gjenbrukes uten gode metadata. Rousidis et al. (2014) påpeker at forskningen så langt har konsentrert seg om kvaliteten på forskningsdataene, eller om metadatakvalitet i digitale arkiver/bibliotek, mens man vet lite om

[4] Hentet fra https://en.uit.no/ub/forskningsstotte/art?p_document_id=473658

metadata i forskningsdata-arkivene (s. 280). Balatsoukas et al. (2018) slår også fast at vi vet lite om bruken av metadata som beskriver forskningsdata og om kvalitetsproblemene knyttet til dette (s. 2). Det er viktig at metadataene er fullstendige og nøyaktige, slik at de kan legge til rette for deling og gjenbruk av forskningsdataene (Rousidis et al., 2014, s. 279). Farnel og Shiri (2014) hevder at bibliotekarene har et bevisst forhold til nødvendigheten av høy kvalitet på metadata for forskningsdata, men at forskerne, som er de som i første omgang lager metadataene for datasettene sine, trenger en økende bevissthet om hvor viktig gode metadata er for gjenfinning, bevaring og gjenbruk av dataene (s. 75).

På nettsiden til The Agricultural Information Management Standards Portal (AIMS)[5] slås det fast at metadata "is arguably one of the most powerful tools available in scholarly communications." Dette viser igjen hvor viktige metadata er for at forskningsdata skal gjøres tilgjengelige for deling og gjenbruk.

Det er derfor på sin plass å sette fokus på dette området, noe Kunnskapsdepartementet (2017) også oppfordrer til i dokumentet "Nasjonal strategi for tilgjengeliggjøring og deling av forskningsdata". Mitt fokus er å undersøke kvaliteten på metadata for forskningsdata som er generert ved *norske* vitenskapelige institusjoner, hvilket det ikke finnes noen forskning på så langt. Dette vil jeg gjøre ved å innhente et datasett med nyere metadata for forskningsdata fra et av de fem generiske arkivene som finnes i Norge.

1.1 Problemstilling

Det finnes ingen forskning om kvaliteten på *norske* metadata for forskningsdata. Det er derfor interessant å undersøke dette feltet, og på denne bakgrunnen, samt begrunnelsene nevnt ovenfor om et generelt behov for å kvalitetssikre metadata for forskningsdata, har jeg formulert følgende problemstilling:

Hva kjennetegner kvaliteten på norske metadata for forskningsdata?

For å svare på dette spørsmålet, velger jeg å bruke FAIR-prinsippene, som sier at forskningsdata og metadata skal være "Findable, Accessable, Interoperable and Reusable" (Wilkinson et al., 2016), som et mål på kvalitet, sammen med kvalitetskrav til metadata uttrykt i tidligere forskning. FAIR-prinsippene ble utviklet for å styrke maskinenes evne til å

[5] http://aims.fao.org/

automatisk finne og bruke forskningsdata, samt styrke menneskers mulighet for å gjenbruke dem (Wilkinson et al., 2016, s. 1), og de har et sterkt fokus på nødvendige egenskaper metadataene må inneha for at dette skal være mulig.

Med utgangspunkt i FAIR-prinsippene har jeg derfor utformet følgende forskningsspørsmål:

F1. I hvilken grad fremmer metadataene gjenfinnbarhet?
F2. I hvilken grad fremmer metadataene interoperabilitet?
F3. I hvilken grad fremmer metadataene gjenbruk?

1.2 Avgrensninger

Som det går frem av mine forskningsspørsmål, har jeg utelatt Accessible (tilgjengelig) fra FAIR-prinsippene. Årsaken er at tilgang har med kommunikasjonsprotokoller å gjøre – at de skal være åpne, standardiserte og tilgjengelig globalt, og dette berører ikke selve metadata-kvaliteten (Wilkinson et al., 2016).

Formalisert lagring og deling av forskningsdata er et relativt nytt felt i Norge. Fordi feltet er i utvikling, må metadatakvaliteten ventes å være stigende ettersom tiden går. Marc et al. (2016) påpekte dette da de avdekket at eldre metadata hadde lavere grad av fullstendighet, nøyaktighet og overensstemmelse. Det er derfor ønskelig å undersøke så nye data som mulig, samtidig som jeg ønsker et størst mulig datasett. Jeg velger derfor å analysere metadata publisert i perioden 2018 – 23.03.2020. Dette vil bli grundigere gjort rede for i metode-kapittelet.

Det finnes flere ulike forskningsdata-arkiver i Norge, og jeg velger å hente mitt datasett fra det arkivet som lagret den største mengden forskningsdata i 2018-2020, nemlig DataverseNO. Ved valg av kilde for innhenting av data har jeg også tatt metodiske hensyn, ved å hente data fra et arkiv som gir meg mulighetene for å analysere metadataenes enkelte felter på en praktisk gjennomførbar måte. Dette vil bli grundig gjennomgått i metode-kapittelet. Jeg velger med andre ord å bruke én av flere mulige kilder for dataene mine, og må ta forbehold om at resultatene kunne ha blitt annerledes hvis jeg hadde brukt andre kilder.

1.3 FAIR-prinsippene

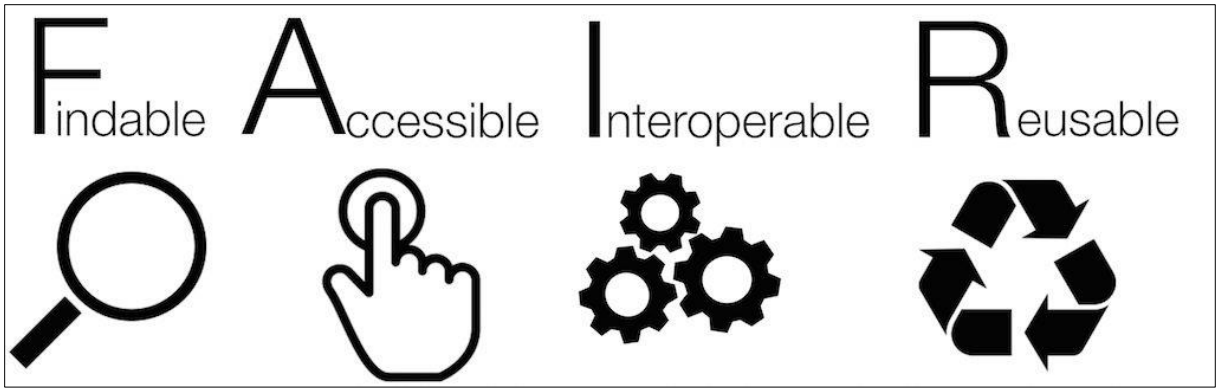
Det enkelte elementet i FAIR-prinsippene vil bli gjort grundigere greie for i teorikapittelet, da de vil utgjøre en vesentlig del av mitt analyseverktøy. Historikken og grunntanken bak FAIR introduseres imidlertid allerede her, fordi de danner grunnlaget for oppgavens forskningsspørsmål.

FAIR omhandler mange aspekter ved infrastrukturen for forskningsdata, inkludert kvaliteten på metadata. I motsetning til tidligere forsøk på å lage retningslinjer for behandling av forskningsdata, utmerker FAIR seg ved å beskrive konkrete, disiplin-uavhengige overordnede prinsipper som kan anvendes bredt i de vitenskapelige miljøene (Wilkinson et al., 2016, s. 4).

På bakgrunn av et sterkt behov for å forbedre infrastrukturen for deling og gjenbruk av forskningsdata, ble "The FAIR Guiding Principles" utformet. Dette skjedde gjennom et samarbeid mellom akademia, industrien, finansierende institusjoner og forskere (Wilkinson et al., 2016). I årene før prinsippene ble utformet, ble det jobbet med prinsipper og retningslinjer som kan sees på som forløpere til FAIR. En publikasjon som ble utgitt som en del av dette arbeidet, var "Principles and Guidelines for Access to Research Data from Public Funding" (OECD, 2007). Man erkjente her at det ikke er tilstrekkelig at forskningsdataene er åpne, men at de også må være tilgjengelige, interoperatible og gjenbrukbare (Hodson et al., 2018, s. 18). Videre kom ministrene i G8-møtet (G8 Science Ministers, 2013) med følgende uttalelse som et år senere ble implementert i FAIR-prinsippene:

> Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards.

FAIR er et akronym bestående av begrepene Findable, Accessible, Interoperable og Reusable, og beskriver hvilke egenskaper både forskningsdata og metadata må ha for at forskningsdata skal få størst mulig verdi og være gjenfinnbare og gjenbrukbare av både mennesker og maskiner (Hodson et al., 2018, s. 10). Figur 2 viser en enkel illustrasjon av prinsippene.



Figur 2: The FAIR Guiding Principles (https://www.panosc.eu/)

Ett aspekt ved FAIR som gjør at de har blitt de rådende prinsippene internasjonalt for håndtering av forskningsdata, er ordspillet som akronymet legger til rette for. Det engelske ordet "fair" assosieres med idealer som likhet og rettferdighet, og har vært uttrykksfullt når man skal kommunisere idéen om at FAIR data tjener både forskerfellesskapenes interesser og forskningens fremme som samfunnstjenlig (Hodson et al., 2018, s. 18).

Behovet for disse prinsippene vokste fram på grunn av de mange generiske forskningsdata-arkivene som etter hvert oppsto. DataverseNO[6], hvor jeg høster mine data, er et eksempel på et slikt arkiv i Norge. I følge Wilkinson et al. (2016) aksepterte disse arkivene en rekke ulike datatyper og formater, gjorde ingen forsøk på å integrere eller harmonisere dataene og stilte få krav med hensyn til beskrivelsen (metadata) av disse dataene. Resultatet ble et data-økosystem som var mindre sentralisert og integrert og mer mangfoldig, noe som forverret muligheten for gjenbruk både for mennesker og maskiner (Wilkinson et al., 2016, s. 2). Willis et al. (2012) beskriver dette som datasiloer, der hvert forskerfellesskap fungerer som en lukket enhet, med egne standarder som forhindrer interoperabilitet (s. 1508). Det var derfor påkrevd å gjøre noe med situasjonen, slik at verdifulle forskningsdata kunne gjenfinnes, repliseres og gjenbrukes.

Wilkinson et al. oppsummerer hva det innebærer at forskningsdata blir behandlet i linje med FAIR-prinsippene (2016):

[6] DataverseNO: https://dataverse.no/Rousidis et al. (2014)

> The goal is for scholarly digital objects of all kinds to become ‘first class citizens’ in the scientific publication ecosystem, where the quality of the publication—and more importantly, the impact of the publication—is a function of its ability to be accurately and appropriately found, reused, and cited over time, by all stakeholders, both human and mechanical. (s. 3)

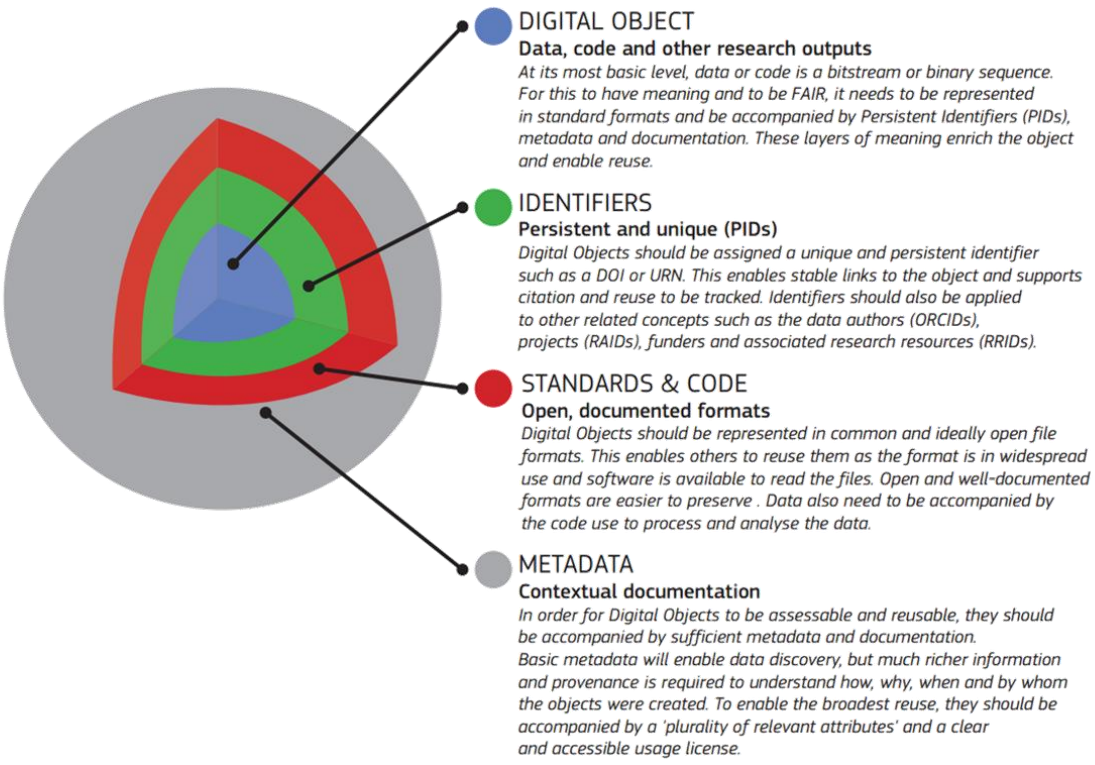
Forfatterne påpeker her at målet for den digitale ressursen er å bli en "first class citizen", med en kvalitet og innflytelse som avhenger av at den kan gjenfinnes, gjenbrukes og siteres over tid av alle interessenter. Dessuten skal dette kunne gjøres av både mennesker og maskiner. Tidligere forsøk på å utforme retningslinjer for håndtering av forskningsdata har hatt fokus på forskeren, mens FAIR vektlegger å styrke maskinenes evne til automatisk å finne og bruke dataene, samt å støtte menneskers gjenbruk av dem (Wilkinson et al., 2016, s. 1). Dette bekreftes også i pressemeldingen som kom etter at FAIR-prinsippene ble publisert: “The recognition that computers must be capable of accessing a data publication autonomously, unaided by their human operators, is core to the FAIR Principles. Computers are now an inseparable companion in every research endeavour” (Mons et al., 2017, s. 51). Her blir det slått fast at selve kjernen i FAIR-prinsippene er datamaskinene og en best mulig tilretteleggelse for at disse kan gjøre forskningsdata gjenfinnbare, tilgjengelige, interoperatible og gjenbrukbare.

Mennesker og maskiner møter ulike utfordringer når det gjelder håndtering av data. I motsetning til maskinene, har vi mennesker en intuitiv forståelse for semantikk, det vil si meningen eller hensikten med et digitalt objekt, ut ifra vår tolkning av kontekst i form av for eksempel tekst eller layout. Våre begrensninger som mennesker oppstår i møte med store mengder data som skal prosesseres, hastigheten som behøves og den tekniske kompleksiteten. Målet med FAIR-prinsippene er at maskinen, i møte med et hittil ukjent digitalt objekt, skal kunne a) identifisere typen objekt (struktur og hensikt), b) avgjøre om objektet stemmer overens med brukerens behov ved å spørre metadata og/eller data-elementer, c) avgjøre om objektet er gjenbrukbart, med hensyn til lisens, godkjenning eller andre begrensinger, og d) handle deretter, på samme måte som et menneske ville ha gjort det (Wilkinson et al., 2016, s. 3).

Fordi maskinene er i fokus, innebærer det at metadata er viktige, da det er de som legger til rette for at maskinene "forstår" dataene. Wilkinson et al. (2016) utdyper at "ultimate machine-actionability occurs when a machine can make a useful decision regarding data that it has not encountered before" (s. 4). Målet er med andre ord at maskinen skal kunne ta beslutninger på bakgrunn av de data og metadata den møter. At maskinen forstår dataene, gjelder for det første de kontekstuelle metadataene som omgir et digitalt objekt ('hva er det?'), dernest innholdet i det digitale objektet ('hvordan prosesserer/integrerer jeg det?'). Disse to prosessene handler om henholdsvis "machine-actionable" metadata og "machine-actionable" data, og er kjernen i det FAIR-prinsippene tilstreber å oppnå (Wilkinson et al., 2016, s. 4).

For at en fil med lagrede forskningsdata skal kvalifisere som et "FAIR digital object", må den, ifølge Hodson et al. (2018) oppfylle følgende kriterier: Ha blitt tildelt en unik, vedvarende identifikator, slik at den kan gjenfinnes, brukes og siteres ; Være lagret i vanlige, og helst åpne, formater ; Være rikelig dokumentert ved hjelp av metadata-standarder og vokabularer som brukes i forskerfellesskapene, og slik legge til rette for interoperabilitet og gjenbruk ; og Inkludere dokumentasjon som gir teknisk veiledning om gjenbruk, samt informasjon om lisenser (s. 12).

Videre illustrerer Hodson et al. (2018) et "FAIR digital object" som vist i Figur 4 nedenfor. I kjernen av objektet finnes selve forskningsdataene, lagret i et standardisert og tilgjengelig format. Dernest finner vi de varige identifikatorene, som for eksempel DOI og ORCID, som er nødvendige for gjenfinning og sitering. Disse er så omgitt av standarder og koder, det vil si at dataene er lagret i åpne filformater som legger til rette for gjenbruk. Ytterst på figuren finner vi metadata, det vil si dokumentasjon som med sin rikholdige og standardiserte informasjon legger til rette for gjenfinning, vurdering av objektet og eventuell gjenbruk.



Figur 3: "A model for FAIR Digital Objects" (Hodson et al., 2018, s. 35).

1.4 Oppbygging av oppgaven

Jeg vil videre gjøre rede for mine litteratursøk, for så å presentere oppgavens kilder.

I teori-kapittelet blir sentrale begreper som er en del av oppgavens problemstilling og analyseverktøy definert og diskutert. Deretter blir metoden for innhøsting og bearbeiding av datasettet som er gjenstand for denne undersøkelsen gjort rede for.

Kvalitetskriterier i tidligere forskning blir så, sammen med FAIR-prinsippene, formet til et analyserverktøy som jeg kan bruke for å vurdere kvaliteten på metadataene i datasettet mitt, og dette analyseverktøyet utgjør undersøkelsesmetoden.

Videre presenterer jeg resultatene fra undersøkelsen, for deretter å drøfte disse opp mot mitt analyseverktøy og oppgavens forskningsspørsmål og problemstilling. Avslutningsvis konkluderer jeg ut ifra mine funn, vurderer undersøkelsens validitet og analyseverktøyets egnethet, samt antyder nytten av denne undersøkelsen og hvordan den kan anspore til videre forskning.

2 Tidligere forskning

Jeg vil her gjøre rede for hvordan jeg har søkt for å finne relevant forskning som angår oppgavens problemstilling. Deretter blir alle oppgavens kilder introdusert.

2.1 Litteratursøk

Jeg begynte litteratursøket i august 2019, og søkte i Oria og Google Scholar, i første omgang på emneord, men etter hvert også på siteringer av aktuelle artikler. Litteraturlistene i valgte artikler er også blitt studert, og det ga meg en nyttig oversikt over litteraturen, da mange av de samme referansene gikk igjen i de ulike litteraturlistene. Dette er ikke overraskende, da forskningsfeltet er relativt nytt og litteraturen på området derfor begrenset.

Jeg søkte først med emneordene: "research data metadata quality". Dette ga meg et viktig treff, som igjen, via litteraturliste og siteringer, ledet meg videre til annen relevant litteratur. Videre søkte jeg med emneordene "scientific data metadata" og fikk ett relevant treff i hver av basene. Deretter søkte jeg på termer som "metadata schema", "metadata record" og "data repository", kombinert med "research data", men det meste av litteraturen fant jeg via siteringer og litteraturlister.

2.2 Litteratur

I det følgende presenteres alle oppgavens kilder; først oversiktslitteratur og kilder som har bidratt til teorier, definisjoner og forståelse av begreper, dernest tidligere forskning om kvaliteten på metadata for forskningsdata, samt forskning om kontrollerte vokabularer i forskningsdata-arkiver.

2.2.1 FAIR-prinsippene

Kunnskapsdepartementet (2017) har publisert rapporten "Nasjonal strategi for tilgjengeliggjøring og deling av forskningsdata", der de presenterer regjeringens krav og forventninger til forskningsmiljøene når det gjelder tilgjengeliggjøring og gjenbruk av offentlig finansierte forskningsdata. Her fremholdes FAIR-prinsippene som rådende og anbefalte retningslinjer, og viktigheten av høy kvalitet på metadata påpekes (s. 25). Wilkinson et al. (2016) er den mest sentrale kilden som omtaler FAIR i detalj, samt historien bak prinsippenes tilblivelse. Noen av forfatterne bak denne artikkelen innså etter hvert at det

oppsto ulike tolkninger av hva "FAIRness" innebærer i praksis, og publiserte derfor artikkelen "Cloudy, increasingly FAIR" i 2017 (Mons et al.). Til hjelp for bedre å forstå hva det enkelte prinsippet innebærer når det gjelder metadata spesifikt, er nettsiden GO FAIR (u.å.) svært opplysende. Videre har European Commission publisert rapporten "Turning FAIR into reality" (Hodson et al., 2018) som også bidrar til en dypere forståelse av FAIR-prinsippene.

2.2.2 Infrastrukturen for forskningsdata

I publikasjonen "Digital Curation Manual" gjør Day (2010) rede for infrastrukturen for forskningsdata, og særlig metadataenes rolle i kurateringen, på et teoretisk plan, mens Austin et al. (2016) har en praktisk tilnærming til infrastrukturen i sin undersøkelse av ulike forskningsdata-arkiver. Nettsiden til University of Vienna (u.å.) gir en grundig beskrivelse av forskningsdata-kretsløpet, fra innsamling av data til gjenbruk av dem, mens UIO (2015) forklarer kuratorens oppgaver og funksjon. Hva en datahåndteringsplan (DMP) inneholder blir beskrevet på nettsidene til Norsk senter for forskningsdata (NSD, u.å.).

2.2.3 Definisjoner og teori

Definisjonen av begrepet "datakvalitet" er hentet fra Gordon (2013) i hans bok om dataforvaltning, mens "forskningsdata" defineres i en publikasjon fra OECD (2007) som omhandler prinsipper og retningslinjer for deling av offentlig finansierte forskningsdata. Denne publikasjonen omtales av Wilkinson et al. (2016) som en forløper til FAIR-prinsippene.

For å legge et teoretisk fundament for analysen av kvaliteten på metadata for forskningsdata, begynner jeg med å studere litteratur om metadata generelt. Gilliand (2008) presenterer en enkel definisjon av begrepet i et bokkapittelet som gir en grunnleggende innføring i temaet metadata. Caplan (2003) definerer også metadata, i den omfattende boken "Metadata fundamentals for all librarians". Dessuten gir hun en grundig innføring i de tre metadata-typene deskriptive, administrative og strukturelle metadata. Riley (2017) er også en sentral kilde som bidrar til forståelsen av disse tre metadata-typene, mens Haslhofer og Klas (2010) introduserer tre metadata-byggesteiner; metadata instance, metadata schema og schema definition language.

Teori om metadata for forskningsdata finner jeg hos Kim et al. (2019) i deres undersøkelse av 20 ulike skjemaer for deponering av forskningsdata. Hider (2018) gir en grundig innsikt i ulike metadata-skjemaer/-standarder, spesielt Dublin Core, mens nettsiden DataCite Metadata Working Group (2019) gir en detaljert oversikt over skjemaet DataCite. I analysen av metadata-målene for forskningsdata-arkivet Dryad, som Willis et al. (2012) presenterer, gis det også en grundig oversikt over Dryads metadata-skjema, samtidig som en rekke ulike metadata-skjemaer for forskningsdata også blir analysert.

2.2.4 Metadata-kvalitet generelt

En rekke kilder berører teori om kvaliteten på metadata generelt. Gjennom sin undersøkelse av metadata i to digitale arkiver, avdekket Barton et al. (2003) behovet for kvalitetssikring av metadataene. De fant særlig svakheter knyttet til personnavn, emneord, stavefeil og forkortelser. Yasser (2011) hadde også fokus på problemområder i forbindelse med kvaliteten på metadata, og fant de fem kategoriene *1) incorrect values; (2) incorrect elements; (3) missing information; (4) information loss; og (5) inconsistent values* (s. 60). Et mye omtalt rammeverk for evaluering av metadata-kvalitet ble utviklet av Bruce og Hillmann (2004). Dette ble utformet på grunnlag av de syv karakteristikkene *Completeness, Provenance, Accuracy, Conformance to expectations, Logical consistency and coherence, Timliness og Accessibility* (s. 243-248). Både Tani et al. (2013) og Ochoa (2014) studerte og evaluerte ulike rammeverk for metadata-kvalitet, og konkluderte begge med at Bruce og Hillmann sitt rammeverk var det mest brukbare.

2.2.5 Interoperabilitet

Interoperabilitet blir definert av NISO (2004), og videre gjort grundig rede for på et teoretisk plan av Haslhofer og Klas (2010). De omtaler fire nivåer av interoperabilitet; teknisk, syntaktisk, semantisk og organisatorisk. Lewis et al. (2008) går enda dypere inn i disse fire nivåene. Det finnes også andre typer interoperabilitet enn den som er knyttet til metadata, og Miller (2000) nevner "political/human, intercommunity, legal, and international interoperability". I forbindelse med rettigheten til gjenbruk av forskningsdata er særlig den legale interoperabiliteten interessant, noe RDA-Codata Legal Interoperability Interest Group (2016) belyser i detalj.

2.2.6 Kvaliteten på metadata for forskningsdata

Andelen forskningsartikler som går direkte på analyse av metadatakvalitet for forskningsdata er meget begrenset. Rousidis et al. (2014) foretok en innledende/forberedende undersøkelse av forskningsdata-arkivet The Dryad repository for å kartlegge bruken av de ulike metadata-feltene og hovedutfordringene knyttet til kvaliteten på metadataene. De undersøkte spesielt utfyllingen av feltene *Creator*, *Date* og *Type*, og fant mange uregelmessigheter på grunn av manglende standardisering. Videre foretok det samme forskerfellesskapet en ny undersøkelse av samme arkiv, der fokus var kvaliteten på utfylling av feltet *Subject* (Rousidis et al., 2015). Funnene var også her metadata med dårlig kvalitet på grunn av manglende standardisering og kontrollert vokabular. Denne undersøkelsen er også presentert i en nyere artikkel skrevet av noen av de samme forfatterne (Balatsoukas et al., 2018). Farnel og Shiri (2014) analyserte fire forskningsdata-arkiver for å kartlegge hvilke metadata-felter som ble brukt, samt hvordan de ble brukt. De fokuserte blant annet på kontrollerte vokabularer og behovet for å tilrettelegge for interoperabilitet. Marc et al. (2016) analyserte metadatakvaliteten i et forskningsdata-arkiv for helsedata, og fant blant annet at metadata lagret i tidligere år hadde lavere grad av fullstendighet, nøyaktighet og overensstemmelse.

Andre forskere har ikke hatt kvaliteten på metadata for forskningsdata som fokus, men har allikevel berørt temaet. Kim et al. (2019) undersøkte hvilken informasjon forskningsdata-arkiver for tre ulike disipliner etterspurte ved deponering av forskningsdata. I en oversikt over tidligere studier, viste de til sviktende metadata-kvalitet ved utfyllingen av enkelte felter: Feltet *Creator* var ufullstendig utfylt; feltet *Date* var inkonsekvent utfylt med varierende formater; feltet *(Resource)Type* var inkonsekvent utfylt; og feltet *Subject* hadde manglende kontrollert vokabular. De fant også at tidligere forskning anbefalte følgende tiltak for å sikre god metadatakvalitet for forskningsdata: unik forfatter-ID, standardisert formattering og lister med faste verdier ved utfylling av metadata-feltene (s. 847).

I en rapport om kuratering av forskningsdata, nevner Choudhury et al. (2018) en rekke krav til forskningsdata-spesifikke metadatafelt. Disse kan sammenfattes som (1) viktigheten av en grundig beskrivelse av datasettet og metoden de ble utvunnet ved hjelp av, hvilket først og fremst tilsvarer den informasjonen man finner i metadatafeltet *Description*, samt (2) nødvendigheten av å oppgi informasjon om brukerrettigheter, som feltet *Lisence* gir informasjon om.

2.2.7 Kontrollerte vokabularer for forskningsdata

På bakgrunn av at det kun finnes begrenset kunnskap om utbredelsen av kontrollerte vokabularer i forskningsdata-arkiver, gjennomførte Zhang et al. (2015) en survey blant ulike aktører innen forskningsdata-forvaltning, og fant ut at deltakerne så på kontrollerte vokabularer som verdifulle hjelpemidler. Dessuten kom det frem et ønske om en teknologi som la til rette for at forskningsdata-arkivene kunne tilby tilgang til flere ulike vokabularer. Karimova (2018) beskriver sitt forestående prosjekt, der målet er å utvikle metadata-modeller med tilhørende kontrollerte vokabularer, fortrinnsvis fleksible skjema som kan brukes sammen med vokabularer fra ulike disipliner. Hun argumenterer for at kontrollerte vokabularer knyttet til ulike metadata-felt vil gjøre deponeringen av forskningsdata enklere for forskeren, slik at han/hun motiveres til å dele sine data. Samtidig øker metadata-kvaliteten og tilrettelegger for interoperabilitet og gjenbruk.

3 Teori

I det følgende vil sentrale begreper som angår oppgavens problemstilling defineres og forklares, og grunnlaget for forskningsspørsmålene – FAIR-prinsippene – vil bli gjort grundig rede for.

Det er nærliggende å sammenligne arkiver for forskningsdata med digitale bibliotek, da begge deler handler om lagring av digitale informasjonsobjekter. En rekke studier er blitt utført på metadata-kvaliteten i digitale samlinger, og Ochoa (2014) gir en oversikt over de mest betydningsfulle. Da denne oppgavens mål er å undersøke kvaliteten på metadata for en helt spesiell type digitale objekter, nemlig forskningsdata, er det imidlertid ikke tilstrekkelig å bruke kvalitetskriteriene i de nevnte studiene som eneste parametere. Jeg velger derfor å ta utgangspunkt i FAIR-prinsippene (Wilkinson et al., 2016), som også er bakgrunnen for oppgavens forskningsspørsmål. Disse er spesielt utviklet for å forbedre infrastrukturen for forskningsdata, og stiller spesifikke krav som metadataene for forskningsdata må innfri for å legge til rette for gjenfinning og gjenbruk av forskningsdata. Disse kravene velger jeg å bruke som mål på kvaliteten på metadata for forskningsdata.

Slik jeg ser det, er det imidlertid ikke tilstrekkelig å kun se til FAIR-prinsippene for å svare på oppgavens problemstilling og forskningsspørsmål. Jeg trenger også definisjoner og forståelser av sentrale begreper som inngår i problemstilling og forskningsspørsmål, og sammen med FAIR-prinsippene kan disse bidra til å forme et analyseverktøy.

Først vil jeg gå til litteraturen for å finne definisjoner på begrepet *metadata* generelt, og dernest spesifikt *metadata for forskningsdata*. Dette for bedre å forstå det fenomenet jeg skal analysere, samt legge grunnlaget for utvelgelsen av hvilke metadata-felt som egner seg for analyse i min undersøkelse. Som et viktig begrep i FAIR-prinsippene og ett av forskningsspørsmålene blir så *interoperabilitet* definert og diskutert. Videre vil jeg gjøre greie for hvordan konseptet *metadata-kvalitet* er blitt forsøkt definert og modellert av ulike forskere. Jeg undersøker så hvordan *kvaliteten på metadata for forskningsdata* er blitt analysert i tidligere forskning, for å finne bidrag til utformingen av mitt analyseverktøy. Deretter setter jeg meg grundig inn i FAIR-prinsippene, og gjør rede for hvilke av de kravene de fremmer som omhandler kvaliteten på metadata.

3.1 Metadata

Begrepet *metadata* kommer opprinnelig fra datavitenskapen. Forstavelsen "meta" betyr "om", hvilket definerer metadata som "data som beskriver andre data" (Caplan, 2003, s. 1), eller kort og godt "data om data" (Gilliand, 2008, s. 1). Det finnes en rekke ulike definisjoner og forståelser av begrepet, avhengig av den sammenhengen det blir brukt i. Noen begrenser begrepet til kun å gjelde digitale metadata, eller metadata om digitale objekter, mens andre har en mer generell forståelse av begrepet. W3C[7] definerte metadata som "machine understandable information for the web" (sitert av Caplan, 2003, s. 2) og ekskluderte på den måten både papirbaserte kataloger på kort eller i bøker, og ikke-web-baserte digitale datasystemer. Den andre ytterligheten er Assosiation of American Publishers definisjon: "Metadata is information that describes content. An everyday example is a card catalog in a library, an entry in a book catalog, or the information in an online index" (sitert av Caplan, 2003, s. 2). Dette er en meget vid definisjon, som ikke beskriver hvordan informasjonen presenteres, og som inkluderer alle medier. Som en gyllen middelvei definerer Caplan (2003) metadata som "structured information about an information resource of any media type or format" (s. 3). Denne definisjonen påpeker at informasjonen må være strukturert, men tar verken hensyn til om informasjonen (metadata) eller informasjons-ressursen som beskrives er digital eller ikke, eller nettbasert eller ikke. Det viktige er to vesentlige aspekter ved metadata: De er strukturerte, og de beskriver en informasjons-ressurs. At metadata er strukturerte, betyr at de ikke består av en tilfeldig samling av data-elementer, men at disse er lagret i et dokumentert metadata-skjema (Caplan, 2003, s. 3).

I konteksten for denne oppgaven, nemlig kvaliteten på metadata for forskningsdata som er lagret og skal gjenfinnes digitalt, må begrepet metadata defineres mer begrenset. Haslhofer og Klas (2010) innleder sin artikkel med å hevde at "metadata is machine processable data that describes resources, digital or non-digital" (s. 1). Her slås det fast at metadata behandles maskinelt, men at ressursene beskrevet kan være digitale eller ikke-digitale. Enda mer presist for denne oppgavens fokus blir det når de videre definerer "metadata as the sum total of what we can say about any information object at any level of aggregation, in a machine understandable representation" (s. 5). Metadata innebærer her med andre ord alt som kan sies om et informasjons-objekt, uansett hvordan det er innhentet, og denne informasjonen må presenteres og kunne forstås maskinelt.

[7] The World Wide Web Consortium, https://www.w3.org/

Videre skal metadata "fulfil a useful purpose" (Riley, 2017, s. 4). Farnel og Shiri (2014) definerer metadata som strukturert informasjon som gir en kontekst for alle slags informasjonsobjekter, inkludert forskningsdata, og som slik legger til rette for bruk, bevaring og gjenbruk av disse objektene (s. 75). Denne definisjonen uttrykker viktigheten av strukturert informasjon (metadata) som en forutsetning for bruk, bevaring og gjenbruk av forskningsdata, og implisitt viktigheten av at kvaliteten på disse metadataene er så høy at de oppfyller sin hensikt.

For bedre å forstå alle aspektene ved metadata, er det vanlig å klassifisere begrepet i tre ulike *typer* som beskriver metadataenes funksjon; nemlig *deskriptive, administrative* og *strukturelle* metadata (Caplan, 2003; Haslhofer & Klas, 2010; Riley, 2017). I følge Haslhofer og Klas (2010, s. 5) er dette en klassifisering som fokuserer på den funksjonen metadataene skal støtte.

Deskriptive metadata gir informasjon om en ressurs for at den skal kunne gjenfinnes og forstås (Riley, 2017, s. [10]). Typiske eksempler er informasjonen om forfatter, tittel og emneord.

Administrative metadata gir informasjon som legger til rette for forvaltning av ressursen som beskrives, og informasjon om ressursens opprinnelse. De kan deles opp i tre underkategorier: *Technical metadata, preservation metadata* og *rights metadata*. *Technical metadata* gir informasjon om hvordan filen beskrevet kan dekodes og gjenbrukes; *preservation metadata* gir informasjon om hvordan filen skal forvaltes over tid; og *rights metadata* gir informasjon om brukerrettighetene knyttet til filen (Riley, 2017, s. [10]).

Den tredje typen metadata, *structural metadata*, er metadata av en helt annen karakter enn de deskriptive og administrative. Strukturelle metadata har å gjøre med de enkelte elementers relasjon til hverandre, dvs. hvordan ressursen er organisert internt. De er viktige for forvaltning og bevaring av ressursen, men også for fremstillingen av den (Caplan, 2003, s. 158).

Riley (2017) nevner også *markup languages* som en fjerde kategori (s. [10]). Disse språkene setter sammen metadata og ressursens innhold for å få frem strukturelle og semantiske egenskaper ved innholdet.

Tabellen nedenfor (Tabell 1) gir en oversikt over disse metadata-typene og de funksjonene de oppfyller.

Descriptive metadata	For finding or understanding a resource
Administrative metadata - Technical metadata - Preservation metadata - Rights metadata	 - For decoding and rendering files - Long-term management of files - Intellectual property rights attached to content
Structural metadata	Relationships of parts of resources to one another
Markup languages	Integrates metadata and flags for other structural or semantic features within content

Tabell 1: Metadata-typer (Riley, 2017, s. [10])

For å få en bedre forståelse av innholdet i begrepet *metadata*, er det nyttig å klassifisere det i typene nevnt ovenfor. Slik tydeliggjøres de ulike funksjonene til ulike typer metadata, og at for eksempel deskriptive metadata ikke er den eneste typen som har betydning for forvaltningen av en ressurs. Riley (2017) presenterer også en tabell (Tabell 2) som illustrerer egenskapene til de ulike metadata-typene, med tilhørende eksempler på egenskaper og bruksområder for den enkelte type. Her gis det eksempler på hvilke egenskaper *(properties)* ved metadata som hører innunder de ulike metadata-typene, noe som samsvarer med metadata-felt, samt hvilke hensikter de fremmer *(primary uses).* Tabellen er gjengitt nedenfor, og vil bli henvist til senere i oppgaven i forbindelse med utvelgelsen av metadata-felt som skal analyseres.

Metadata Type	Example Properties	Primary Uses
Descriptive metadata	Title Author Subject Genre Publication date	Discovery Display Interoperability
Technical metadata	File type File size Creation date/time Compression scheme	Interoperability Digital object management Preservation
Preservation metadata	Checksum Preservation event	Interoperability Digital object management Preservation
Rights metadata	Copyright status License terms Rights holder	Interoperability Digital object management
Structural metadata	Sequence Place in hierarchy	Navigation
Markup languages	Paragraph Heading List Name Date	Navigation Interoperability

Tabell 2: Metadata-typer med egenskaper og bruksområder (Riley, 2017, s. [11]).

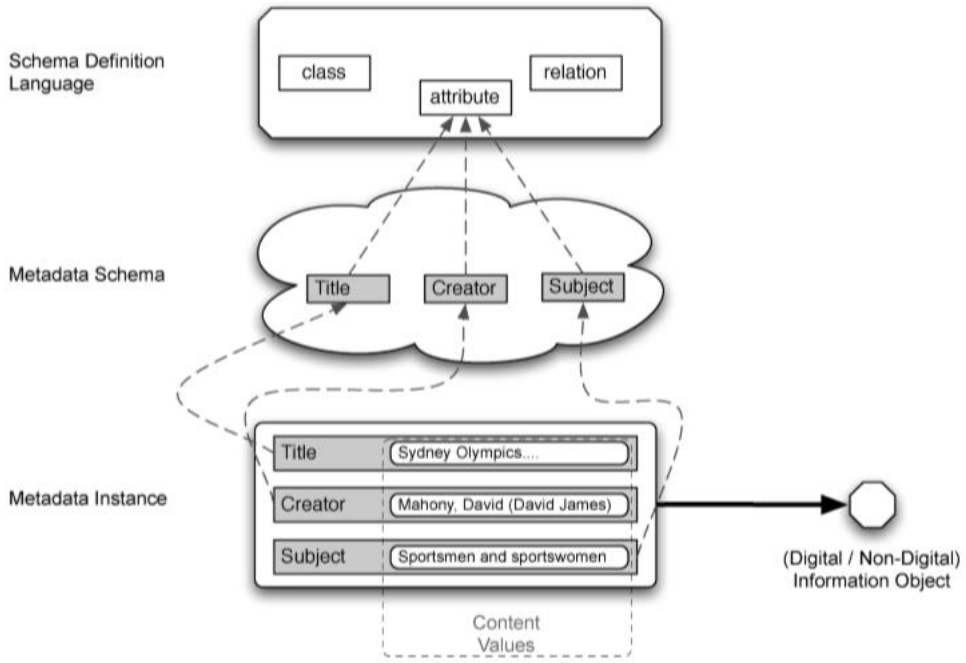
Det finnes også andre måter å klassifisere metadata på. Haslhofer og Klas (2010) deler opp metadata i tre "byggesteiner", og viser hvordan disse sammen utgjør en metadata-post (s. 6-8):

(1) *Metadata instance*: Består av elementer fra et metadata-skjema, sammen med tilhørende innhold/verdier. Denne kombinasjonen av elementer og innhold utgjør metadata-beskrivelsen av et bestemt informasjons-objekt.

(2) *Metadata* schema: Består av de ulike feltene/elementene i metadata-posten.

(3) *Schema Definition Language*: Formatet metadata-posten er lagret i (XML, RDF m.fl.).

Forfatterne har illustrert dette i en modell som gjengis nedenfor (Figur 4).



Figur 4: De tre metadata-byggesteinene (Haslhofer & Klas, 2010).

I min analyse av kvaliteten på metadata for forskningsdata, er det først og fremst byggestein (1) *Metadata Instance* jeg kommer til å ha fokus på, da det er innhold og verdier i metadata-skjemaet som varierer og kan påvirke kvaliteten. Det er imidlertid også aktuelt å ta hensyn til de forutsetninger og eventuelt begrensninger som byggestein (2) *Metadata Schema* impliserer. Dette innebærer både hvordan skjemaet er bygget opp i forhold til hvilke felt som er tilgjengelige, og hvordan det enkelte arkiv legger til rette for registrering av innhold i de enkelte feltene. Det er som oftest på arkiv-nivå bestemmelsen tas om hvorvidt felt skal ha faste verdier eller være fritekst-felt.

UNC University Libraries (2019) definerer metadata-skjema som

> the overall structure for the metadata. It describes how the metadata is set up, and usually addresses standards for common components of metadata like dates, names, and places.

Et metadata-skjema legger til rette for en standardisert og entydig beskrivelse av en ressurs. Det finnes ulike typer skjema, med ulikt antall elementer, obligatoriske felt og standardiserte verdier. Hvilket skjema som brukes vil derfor ha en innvirkning på metadataenes detaljnivå

og kvalitet. Noen skjemaer er generiske, og følgelig ganske generelle, mens skjemaer for bestemte disipliner er mer spesialiserte og bare forståelige for forskerne innenfor den enkelte disiplin (UNC University Libraries, 2019). Dublin Core[8] er det generiske metadata-skjemaet som er mest utbredt, og det er dette jeg vil forholde meg til når jeg analyserer mine data. Andre mye brukte skjemaer er DDI[9] (Data Documentation Initiative) og DataCite[10].

Dublin Core er et enkelt skjema som i utgangspunktet kun inneholder 15 felt. Noen av disse har imidlertid fått underkategorier som kan brukes ved behov (Hider, 2018, s. 127). I tabellen nedenfor (Tabell 3) gjengis de 15 hoved-elementene i Dublin Core.

Identifier	Definition
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content of the resource.
Subject	The topic of the content of the resource.
Description	An account of the content of the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the content of the resource.
Date	A date associated with an event in the life cycle of the resource.
Type	The nature or genre of the content of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A reference to a resource from which the present resource is derived.
Language	A language of the intellectual content of the resource.
Relation	A reference to a related resource.
Coverage	The extent or scope of the content of the resource.
Rights	Information about rights held in and over the resource.

Tabell 3: De 15 elementene i "Simple Dublin Core" (https://www.semanticscholar.org/)

3.2 Metadata for forskningsdata

Metadata for forskningsdata inneholder mye av den samme informasjonen som alle andre metadata for informasjons-ressurser generelt: forfatter, tittel, emneord m.m. Tilrettelegging for lagring, gjenfinning og gjenbruk av forskningsdata stiller imidlertid særskilte krav til metadataene, og derfor inneholder de noen felt som anvendes på en unik måte i arkiver for

[8] https://www.dublincore.org/schemas/
[9] https://ddialliance.org/
[10] https://schema.datacite.org/meta/kernel-4.3/

forskningsdata. Et av disse feltene er *Type*, som brukes til å angi hva slags type data som er lagret, for eksempel "survey data", "clinical data", "machine-readable text" m.fl. Innholdet i dette feltet er av vesentlig betydning for at forskningsdataene skal kunne forstås og gjenbrukes (Rousidis et al., 2014, s. 282).

Choudhury et al. (2018) nevner annen nødvendig informasjon som beskriver forskningsdataene, nemlig innsamlingsmetode, metode for dataprosessering og analyse, samt kontekst (s. 7). Disse opplysningene finnes vanligvis i feltet *Description*. Kim et al. (2019) påpeker også at en grundig beskrivelse av dataenes innhold og kontekst er viktig for gjenbruk, hvilket bekrefter at innholdet i *Description*-feltet er avgjørende for metadata-kvaliteten.

Kim et al. (2019) viser til en klassifisering av metadata for forskningsdata i typene (1) *Intrinsic* og (2) *User-centric* (submitter-defined eller user-expanded) metadata. *Instrinsic metadata* vil si faktisk og udiskuterbar informasjon om det lagrede objektet, som dato for innsamling av data, metadata-skaper (creator), rettigheter m.m. *User-centric metadata* vil si informasjon som er nødvendig for å kunne gjenbruke forskningsdataene, samt en fyldig beskrivelse av konteksten for datainnsamlingen. Denne informasjonen kan legges inn av dataskaperen selv, eller av de som gjenbruker dataene og eventuelt har funnet feil eller bias (s. 863).

En viktig egenskap ved metadata for forskningsdata, er informasjon om brukerrettigheter. Det er viktig at både maskiner og mennesker forstår hvilke rettigheter som er knyttet til de lagrede forskningsdataene, ved at det går klart fram om gjenbruk er tillat (GO FAIR, u.å.). Denne informasjonen legges i feltet *License,* og fylles ut med verdien CC0[11] hvis alle skal ha rettigheter til å gjenbruke forskningsdataene. For at gjenbruk skal være mulig, er det obligatorisk at dette feltet er utfylt med CC0.

3.3 Interoperabilitet

Interoperabilitet er en viktig egenskap ved metadata, da den gjør det mulig for systemer og applikasjoner å jobbe med eller bruke informasjonsobjekter på tvers av systemer (Haslhofer & Klas, 2010, s. 14). NISO (2004) definerer interoperabilitet som "the ability of multiple

[11] https://creativecommons.org/share-your-work/public-domain/cc0/

systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality" (s. 2).

For bedre å forstå interoperabilitet, har flere forskere introdusert en underinndeling av fenomenet, med ulike nivåer av interoperabilitet som kan oppnås i den digitale infrastrukturen (Haslhofer & Klas, 2010; Lewis et al., 2008). Nivå 1 er *Machine level/technical level*. Dette innebærer kommunikasjon og overføring av data mellom ulike typer maskinvare, basert på felles standarder. Nivå 2 er *Syntactic interoperability*, som betyr at ulike programmeringsspråk på ulike plattformer forstår format og struktur på data som utveksles, slik at de kan "snakke" med hverandre. Disse to nivåene er det mulig å oppnå uten for store utfordringer. Det betyr imidlertid ikke at de ulike systemene forstår innholdet i de dataene som utveksles. Nivå 3, *Semantic interoperability*, innebærer at systemene enes om en felles semantikk, det vil si forståelse av ordenes betydning, for alle data som deles. Måten å løse dette på, er inngå uformelle avtaler, gjerne i form av domene-spesifikke ontologier, for å formalisere forståelsen av ord og begreper på tvers av systemer. Dette nivået oppnås så langt bare delvis, og da gjerne ved at mennesker tolker dataene. Nivå 4, *Organizational interoperability*, handler om å ta i betraktning de delte dataenes kontekst når de skal deles på tvers av organisasjoner.

Interoperabilitet innebærer ikke bare at data kan deles mellom systemer, men at man kan gjøre noe med dataene – at de er "actionable". Det vil si at man kan gjøre endringer som forstås av de andre samhandlende systemene, uten at det fører til uønskede konsekvenser for de andre systemene. Mens nivå 3 ofte krever handling fra menneske til data, krever nivå 4 handling mellom mennesker. Dette for å sørge for at informasjon blir behandlet på en måte som tilfredsstiller alle involverte organisasjoner (Lewis et al., 2008, s. 3-4).

Den teknologiske utviklingen støtter så langt maskin-til-maskin transaksjoner på de to første nivåene, og til en viss grad på det tredje nivået. Å oppnå interoperabilitet på nivå 3 og 4 er imidlertid mye mer komplisert og i mange tilfeller ennå ikke mulig (Lewis et al., 2008, s. 4).

Haslhofer og Klas (2010) nevner flere klassifiseringer av interoperabilitet beskrevet i litteraturen, og de fleste har fellestrekk med nivåene nevnt ovenfor. I tillegg trekker de frem Miller (2000) sin inndeling av typer interoperabilitet som ikke angår det tekniske aspektet; nemlig "political/human, intercommunity, legal, and international interoperability." I

konteksten forskningsdata er det interessant å se på konseptet *legal interoperability* (juridisk interoperabilitet). Research Data Alliance har publisert retningslinjene "Legal interoperability for research data" (RDA-Codata Legal Interoperability Interest Group, 2016), fordi de har observert misforståelser og mangel på kunnskap og veiledning når det gjelder det juridiske aspektet ved forskningsdata generelt. Dette feltet handler om opphavsrettigheter til forskningsdataene, der full juridisk interoperabilitet oppnås når det ikke foreligger noen begrensninger med hensyn til gjenbruk av dataene. Nærmere beskrevet innebærer juridisk interoperabilitet at (1) De juridiske betingelsene for bruk for hvert datasett er klare og utvetydige ; (2) De juridiske betingelsene for bruk for hvert datasett tillater gjenbruk og kombinasjon av flere datasett ; (3) Brukerne har juridisk tilgang til hvert datasett uten å måtte søke eieren om tillatelse for hver gang de skal bruke datasettene (RDA-Codata Legal Interoperability Interest Group, 2016, s. 1). Selv om Haslhofer og Klas (2010) betegner juridisk interoperabilitet som noe på siden av det tekniske, har den absolutt et teknisk aspekt. De juridiske betingelsene må nemlig lagres maskinelt i et metadata-felt, noe jeg vil komme tilbake til ved valg av hvilke metadata-felt som skal analyseres i denne undersøkelsen.

Et viktig spørsmål er hva som forhindrer interoperabilitet, og det overordnede svaret er *heterogenitet.* Haslhofer og Klas (2010) deler dette inn i begrepene *structural heterogenity* og *semantic heterogenity*, der det første innebærer ulikheter knyttet til oppbyggingen av metadata-skjema, mens det andre handler om konflikter knyttet til bruk av terminologi (s. 14-16). Videre presenterer de teknikker for å oppnå interoperabilitet, og tar utgangspunkt i inndelingen av de tre metadata-byggesteinene som ble nevnt under delkapittelet om metadata: (1) *Metadata instance*, som er innholdet i metadata-posten; (2) *Metadata schema,* som er de ulike feltene/elementene i metadata-posten; og (3) *Schema Definition Language*, som er formatet metadata-posten er lagret i (XML, RDF m.fl.) (s. 6-8). Ut ifra tabellen deres (s. 17) har jeg hentet ut følgende faktorer som legger til rette for interoperabilitet: På det første nivået, *Metadata instance,* vil kontrollerte vokabularer og autoritetsregistre for personnavn forbedre interoperabiliteten. Zhang et al. (2015) definerer kontrollerte vokabularer som " semantic systems that are useful for organizing and accessing resources — and supporting semantic interoperability among object descriptions and repositories" (s. [1]). Her uttrykkes viktigheten av kontrollerte vokabularer for å oppnå interoperabilitet mellom ulike forskningsarkiver. Videre vil et standardisert metadata-skjema styrke interoperabiliteten på nivået *Metadata schema*, mens et standardisert format vil legge til rette for interoperabilitet på nivået *Schema Definition Language.*

3.4 Kvalitetskrav til metadata generelt

Forskerne er klare på viktigheten av metadata av høy kvalitet, samtidig som det ikke finnes noen enighet om hva metadata-kvalitet er eller hvordan den kan måles (Ochoa, 2014; Tani et al., 2013). Det man har blitt enige om, er at utfordringene med å definere begrepet kommer av dets multidimensjonale og kontekstavhengige karakter (Tani et al., 2013, s. 1195). En reell utfordring med metadata er nettopp konteksten de er en del av. Metadata kan fungere tilfredsstillende innenfor et spesielt arkiv eller en disiplin, men når det kreves interoperabilitet – at data kan bevege seg fritt mellom ulike systemer, kan utfordringer oppstå for eksempel i form av at parametere som garanterte for høy kvalitet i det opprinnelige arkivet gir motsatt effekt i et annet system (Tani et al., 2013, s. 1196).

På samme måte som datakvalitet helt enkelt kan defineres som "fitness for use" (Gordon, 2013), kan metadata-kvalitet også gis denne definisjonen. Og i denne enkle definisjonen ligger det kontekstavhengige – om kvaliteten oppfattes som høy avhenger av hva metadataene skal brukes til, eller i hvilken kontekst de skal fungere. Dette vil jeg komme nærmere inn på når jeg i neste delkapittel skal fokusere spesifikt på forståelsen av konseptet *kvaliteten på metadata for forskningsdata*.

Ingen har klart å definere begrepet *metadata-kvalitet*, men har heller laget rammeverk for å identifisere og vurdere kvalitets-parametere i ulike kontekster (Ochoa, 2014; Tani et al., 2013). Tani et al. (2013) undersøkte hvordan utviklingen har vært i litteraturen når det gjelder forsøk på å definere metadata-kvalitet, og gir en oversikt over syv ulike rammeverk over metadata-kvalitet som ble publisert i årene 2004-2011. Ochoa (2014) har også studert de ulike rammeverkene, og begge slår fast at Bruce og Hillmann (2004) sitt rammeverk er blitt mest referert til av andre forskere innenfor fagfeltet metadata-kvalitet. Det består av følgende kvalitets-egenskaper: "Completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility" (s. 243). Fordi intensjonen er at disse egenskapene skal være uavhengige av disiplin, er de relativt abstrakte (s. 242). Tani et al. (2013) etterlyser imidlertid formelle definisjoner av disse begrepene, samt metriske mål som graderer oppnåelsen av kvalitet innenfor det enkelte begrepet (s. 1196). Dette og de andre rammeverkene som de presenterer, er imidlertid først og fremst utviklet for å beskrive egenskapene ved metadata-kvalitet, og i hvilken grad en ressurs kan ses på som av høy kvalitet, basert på disse beskrivelsene (s. 1199).

Barton et al. (2003) refererte til områder hvor problemene med metadata-kvalitet som oftest oppstår: staving, forkortelser, navneformer, emneord (standardisert terminologi) m.m. (s. [3]). Gjennom litteraturstudier om problemer med metadata-kvalitet, kom Yasser (2011) frem til fem problem-kategorier: *(1) incorrect values; (2) incorrect elements; (3) missing information; (4) information loss; og (5) inconsistent values* (s. 60). På denne måten viser han hva som *ikke* er høy kvalitet, og at det motsatte må etterstrebes for å høyne kvaliteten. Funnene i disse undersøkelsene vil jeg ta med meg når analyseverktøyet mitt skal utformes.

3.5 Kvalitetskrav til metadata for forskningsdata

Forskerne er enige om at kvaliteten på metadata for forskningsdata er et viktig område å undersøke, og at vi foreløpig har liten kunnskap om den (Austin et al., 2016; Balatsoukas et al., 2018). Rousidis et al. (2014) viser til definisjoner på datakvalitet i litteraturen, og at både datakvalitet og metadatakvalitet har de samme egenskapene og de samme utfordringene (s. 280). Det impliserer at når Gordon (2013) definerer datakvalitet som "the state of completeness, validity, consistency, timeliness and accuracy that makes data suitable for a specific use" (s. xviii), så beskriver dette også kvaliteten på metadata for forskningsdata.

Den tidligere nevnte definisjonen av datakvalitet som "fitness for use" (Gordon, 2013) kan utdypes i konteksten kvalitetskrav til metadata for forskningsdata. Det kontekstavhengige – det vil si hva metadataene skal brukes til, eller i hvilken kontekst de skal fungere – har stor betydning når det gjelder forskningsdata. Deres tilhørende metadata må blant annet gi informasjon som tilrettelegger for gjenbruk; det vil si informasjon om datatype, rettigheter og en grundig beskrivelse av dataene og undersøkelsen de er grunnlag for. For at dette skal være mulig, må metadataene også gjøre forskningsdataene gjenfinnbare og interoperatible. Dette er elementer i FAIR-prinsippene som vil belyses grundig i delkapittelet nedenfor.

Videre påpeker Rousidis et al. (2014) at litteraturen viser at kvaliteten på metadata for forskningsdata er disiplinavhengig, noe som gir kvalitet en kulturell dimensjon. Kim et al. (2019) utdyper at bredden og dybden i den beskrivende informasjonen som skal følge forskningsdata-settet, varierer mellom disipliner. Samtidig trekker Rousidis et al. (2014) frem utfordringer knyttet til metadatakvalitet som finnes på tvers av disipliner, som for eksempel feilstaving, ulike versjoner av personnavn, manglende standardisering av emneord m.m. (s.

280). Da jeg skal analysere et datasett hentet fra et generisk arkiv, er dette utfordringer det vil være aktuelt å sette fokus på.

Rousidis et al. (2014) beskriver metadata for forskningsdata av høy kvalitet som *complete* og *accurate,* og påpeker at disse egenskapene er viktige blant annet for å legge til rette for gjenbruk og deling av forskningsdata (s. 279). I vurderingen av metadatakvalitet i et "Health Data Repository" brukte Marc et al. (2016) egenskapene *completeness, accuracy* og *consistency* som kvalitetskriterier (s. 865). Dette samsvarer med funnene Kim et al. (2019) gjorde etter å ha studert tidligere forskning, nemlig at metadata i arkiver for forskningsdatadata ofte er *incomplete, inaccurate* og *inconsistent* i format og terminologi (s. 847). Vi finner også disse kriteriene i Bruce og Hillmann (2004) sin modell for metadatakvalitet generelt*: Completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility* (s. 243).

Ut ifra disse funnene virker det hensiktsmessig å gjøre bruk av kvalitetskriteriene *completeness, accuracy* og *consistency* (heretter: *fullstendighet, nøyaktighet og overensstemmelse*) i min analyse, da det er disse som er brukt i forskning om metadatakvalitet for forskningsdata spesielt. Bruce og Hillmann (2004) forklarer at *fullstendige* metadata bør beskrive informasjons-ressursen så fullstendig som tid og midler tillater det. *Nøyaktighet* innebærer redigering av høy kvalitet, for eksempel ved at stavefeil unngås, standardiserte navneformer og forkortelser benyttes m.m. (s. 243). *Overensstemmelse* innebærer en sammenheng mellom bruk av emneord, definisjoner og konsepter i de ulike metadata-postene. Dette er for eksempel noe brukeren forventer ved søk; at de kan søke etter lignende ressurser i ulike samlinger/arkiver ved hjelp av samme søkekriterier, og at søkeresultatene har tilnærmet lik struktur og utseende (s. 245).

Oppsummert viser forskningen at metadata for forskningsdata må ha egenskapene *fullstendighet, nøyaktighet* og *overensstemmelse,* og at *standardisering* og *kontrollerte vokabularer* er viktige virkemidler for å sikre kvaliteten.

3.6 FAIR-prinsippene i detalj

Sammen med parameterne for metadata-kvalitet nevnt ovenfor, vil FAIR-prinsippene danne grunnlaget for analysen av datasettet. De er også utgangspunktet for oppgavens tre

forskningsspørsmål. Etter å ha fokusert på historikken og tankegangen bak prinsippene tidligere i oppgaven, vil jeg nå gå inn i det enkelte elementet i detalj. Riktignok hevder Hodson et al. (2018) at "notions of findability, accessibility, interoperability and reusability - and the actions needed to enable them - are so deeply intertwined that it does not make sense to address them individually" (s. 10). De fokuserte derfor på kulturen i forskningsmiljøene, samt teknologien som må fungere for at data skal etterkomme FAIR-prinsippene. De gir imidlertid allikevel en oversikt over de fire prinsippene med tilhørende forklaringer, og vil sammen med Wilkinson et al. (2016) og GO FAIR (u.å.) danne grunnlaget for min utgreiing av FAIR-prinsippene.

I tabellen nedenfor (Tabell 4) forklarer Wilkinson et al. (2016) det enkelte prinsippet. I det følgende vil jeg gå igjennom disse i detalj, og da spesifikt de punktene som etter min vurdering angår metadata-kvalitet. Som tidligere nevnt utelater jeg prinsippet *Accessible*, da det handler om kommunikasjonsprotokoller og dermed ikke berører kvaliteten på metadata.

En del av prinsippene angår både forskningsdata og metadata, og blir da angitt som *(meta)data* i tabellen. Det er imidlertid metadata som er fokus for denne oppgaven.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Tabell 4: FAIR-prinsippene (Wilkinson et al., 2016).

Findable:

a) FAIR-prinsippet om at forskningsdata og metadata skal være gjenfinnbare, både av mennesker og maskiner, innebærer at de er blitt tildelt en *globalt unik og bestandig identifikator* (Wilkinson et al., 2016). Dette er uten tvil det viktigste prinsippet, da det er umulig å innfri de andre FAIR-prinsippene uten unike og vedvarende identifikatorer (GO FAIR, u.å.). I følge GO FAIR (u.å.) er man ved å bruke en slik identifikator allerede kommet langt på vei i å publisere FAIR data. Unike identifikatorer er vesentlige for sitering og referanser, samt koblingen mellom datasett og metadata (Hodson et al., 2018, s. 19). Den mest brukte identifikatoren for metadata er en DOI (Digital Object Identifier). Dette er en bestandig identifikator, standardisert av ISO[12], som gir digitale objekter en unik id. Forskere kan også tildeles en unik identifikator, som oftest ved hjelp av en ORCID[13]-id. En ORCID gir et personnavn en standardisert navneform, slik at man unngår at ulike versjoner av navnet er i bruk.

b) Videre innebærer prinsippet om gjenfinnbarhet at forskningsdata er beskrevet ved hjelp av *rikholdige metadata* (Wilkinson et al., 2016). Dette vil si rikelig og omfattende informasjon, blant annet om forskningsdataenes kontekst, kvalitet, kjennetegn og tilstand, lagret på en måte som legger til rette for at datamaskinen er i stand til å sortere og gjenfinne dem. Resonnementet bak dette prinsippet er at dataene skal være gjenfinnbare, selv uten en unik identifikator (GO FAIR, u.å.). Videre har rikholdige metadata *presise* og *relevante* egenskaper (Wilkinson et al., 2016). Dette kan innebære faste vokabularer og bruk av emneord som er relevante i forskningsmiljøene.

Interoperable:

Wilkinson et al. (2016) definerer interoperabilitet som "the ability of data or tools from non-cooperating resources to integrate or work together with minimal effort" (s. 2). Det innebærer at dataene blir beskrevet ved hjelp av vokabularer og standarder som angir den presise betydningen av konsepter og kvaliteter som dataene representerer, samt bruk av et standardisert metadata-skjema (Hodson et al., 2018). For at maskiner skal kunne lese og utveksle hverandres data, er med andre ord bruk av standardisert(e)

[12] International Organization for Standardization
[13] https://orcid.org/

språk/vokabularer av avgjørende betydning (GO FAIR, u.å.). Samtidig skal behovet for å kunne gi en grundig og nyansert beskrivelse av dataene ("semantic richness") ivaretas, og å imøtekomme begge disse kravene er en av utfordringene som det jobbes med for å oppnå FAIR data (Hodson et al., 2018, s. 40).

Reusable:

a) FAIR-prinsippet om at både data og metadata skal være gjenbrukbare, betyr at de er *rikholdig* beskrevet med en rekke *presise og relevante egenskaper* (Wilkinson et al., 2016). Dette har sammenheng med tilrettelegging for gjenfinning, men her er fokuset at brukeren (menneske eller maskin) skal kunne finne ut om dataene er *brukbare* i den gjeldende konteksten. Det er da viktig at metadataene ikke bare muliggjør gjenfinning, men også beskriver den konteksten dataene ble generert i. Til og med informasjon som kan synes irrelevant for dataskaperen bør inkluderes i metadataene, da det ikke er mulig å forutse brukerens identitet og behov (GO FAIR, u.å.).

   Kunnskapsdepartementet (2017) konkretiserer dette ved å nevne eksempler som detaljer knyttet til innsamling, avgrensninger, definisjoner og annen nødvendig informasjon for å muliggjøre gjenbruk som nødvendige faktorer. Videre påpeker de viktigheten av at datasettene "bør følge relevante standarder, med strukturerte vokabularer, identifikatorer og referanser m.m., for at de enklere kan forstås i forskerfellesskapene, håndteres maskinelt og brukes sammen med andre datasett" (s. 25).

b) Videre innebærer prinsippet om tilrettelegging for gjenbruk at både data og metadata inneholder en klar og tilgjengelig lisens for bruk (Wilkinson et al., 2016). Mens punktet ovenfor handler om teknisk interoperabilitet, handler dette punktet om juridisk interoperabilitet (GO FAIR, u.å.). Det er av avgjørende betydning at brukerrettighetene kommer klart frem, slik at de kan forstås både av mennesker og maskiner. Som tidligere nevnt er det ønskelig at delte forskningsdata gis fulle og ubegrensede brukerrettigheter (RDA-Codata Legal Interoperability Interest Group, 2016, s. 1).

c) (Meta)data inneholder klar informasjon om opprinnelse. Dette kan innebære hvordan dataene ble skapt, hvem som samlet dem, hvordan de ble prosessert og hvordan dataskaperen ønsker å bli sitert (GO FAIR, u.å.).

4 Metode

For å svare på oppgavens problemstilling trenger jeg et datasett som jeg kan analysere. I dette kapittelet vil jeg gjøre rede for hvordan jeg går fram for å innhente og bearbeide et representativt utvalg av norske metadata for forskningsdata. Først vil jeg beskrive de mulige kildene, deretter hvordan datasettet blir høstet, og videre hvordan jeg bearbeider dataene og gjør dem klare for analyse. Til slutt vil jeg gjøre rede for hvordan jeg utvikler mitt analyseverktøy.

4.1 Innsamling av data

Jeg forutsetter at kvaliteten på metadata for forskningsdata er stigende, noe Marc et al. (2016) påviste da de avdekket at eldre metadata hadde lavere grad av fullstendighet, nøyaktighet og overensstemmelse. Derfor ønsker jeg å analysere så nye data som mulig, samtidig som jeg trenger en stor nok mengde data for å gi resultatene størst mulig grad av gyldighet. Av den grunn velger jeg å høste data som er publisert i perioden 01.01.2018 til 23.03.2020.

I Norge har vi fem data-arkiver/infrastrukturer for lagring av forskningsdata som er generiske, dvs. tilbyr tjenester på tvers av de fleste fagområdene (Kunnskapsdepartementet, 2017, s. 20):

- UiT Open Research Data (ved UiT) + tilbyr DataverseNO https://dataverse.no/dataverse/uit + https://dataverse.no/
- TSD (Tjenester for sensitive data) (ved UiO) https://www.uio.no/tjenester/it/forskning/sensitiv/
- NORDi (Norwegian Open Research Data Infrastructure) (ved NSD) https://nsd.no/digitalisering/nordi/
- NIRD (National e-Infrastructure for Research Data) (ved UNINETT Sigma 2 AS) https://archive.norstore.no/
- BIBSYS BIRD (ved BIBSYS – nå UNINETT og Handelshøyskolen BI) https://bird.unit.no/

TSD er kun en tjeneste for sensitive data, og NORDi er en tjeneste som er under utvikling. De er derfor uaktuelle som kilder for datasettet mitt. Ved søk på forskningsdata arkivert i tidsrommet 01.01.2018 – 23.03.2020 i de tre andre arkivene, får jeg følgende resultater: 59 treff hos NIRD, 36 treff hos BIRD og 590 treff hos DataverseNO. Jeg velger da den kilden som gir meg flest treff, nemlig DataverseNO. De har et lett tilgjengelig OAI-PMH (The Open

Archives Initiative Protocol for Metadata Harvesting)[14] som gjør høsting av metadata lett tilgjengelig, samt mulighet for å se på hver post i ulike dataformater via søk på nettsiden. På grunn av størst datamengde og lettest tilgang til data, er det derfor naturlig å velge dette arkivet for å høste datasettet mitt. Det er også ønskelig å gjøre denne undersøkelsen på tvers av disipliner, for å få et så bredt grunnlag for analysen som mulig. BIRD dekker kun tre institusjoner, NIRD dekker fire, mens DataverseNO dekker hele ni norske forskningsinstitusjoner.

Metoden jeg bruker for å høste datasettet er relativt lik den som beskrives av Balatsoukas et al. (2018, s. 3-4): Metadataene høstes ved hjelp av et OAI-PMH. Deretter blir de generert i skripteprogrammet Jupyter, slik at kun de feltene jeg ønsker å analysere blir trukket ut og lagret i en CSV-fil. Denne filen blir så lastet opp i Excel, hvor jeg kan sortere dataene i kolonner og bearbeide dem videre ved behov.

4.2 Bearbeiding av datasettet

Etter å ha gjort en første sortering av dataene som besto av 590 poster, oppdager jeg at rensing er nødvendig. En rekke poster er så å si identiske; dvs. den eneste forskjellen er tidsangivelsen i tittelen. Ett eksempel er værdata for et bestemt geografisk område for hver måned i tidsrommet 1993 – 2018. Det er hensiktsløst å beholde alle disse postene, da det ikke finnes forskjeller på utfyllingen av noen av feltene jeg skal analysere. Jeg beholder derfor bare én av disse postene. Etter å ha slettet andre poster som også er så å si identiske, består mitt nye datasett av 131 poster. Mine data gir ikke et fullstendig bilde av kvaliteten på norske metadata for forskningsdata, men utgjør etter min oppfatning et representativt utvalg.

Ved videre analyse av datasettet, viser det seg at filene jeg har høstet ikke gir meg all den informasjon jeg trenger: Det er ikke mulig å se om man har brukt den standardiserte navneformen fra ORCID i feltet *Creator*, og feltene for henholdsvis *Subject* og *Keyword* heter alle *Subject*. Derfor må jeg søke opp hver enkelt post på www.dataverse.no for å få denne informasjonen. Dette medfører at jeg må veksle mellom de to kildene DataverseNO og mitt Excel-dokument, men metoden er gjennomførbar, da datasettet ikke er så stort. En manuell

[14] https://www.openarchives.org/pmh/

analyse-metode er imidlertid ikke optimal, da den er tidkrevende og kan medføre unøyaktigheter.

De undersøkelsene om kvaliteten på metadata for forskningsdata som jeg har studert, brukte alle en automatisk analyse-metode (Balatsoukas et al., 2018; Kim et al., 2019; Rousidis et al., 2014; Rousidis et al., 2015). Deres datasett var imidlertid også mye større enn mitt, slik at en manuell analyse ikke ville ha vært gjennomførbar.

I delkapittelet "Quality Studies in metadata collections" beskriver Ochoa (2014) hvordan kvalitetsstudier av metadata kan deles inn i to grupper; 'manual studies' og 'automated studies'. Han påpeker at 'automated studies' er langt billigere enn de manuelle, men at "their level of 'meaningfulness' is directly related with how comprehensive and complex is the set of calculated metrics" (s. 73). Konklusjonen hans er likevel at den stadig økende størrelsen på forskningsdata-arkivene ikke muliggjør noen annen metode for kvalitetsevaluering enn den automatiserte (s. 78). Videre forklarer han hvordan manuelle studier kan brukes for å analysere deler av en samling, det vil si et statistisk signifikant utvalg av metadatapostene ifølge et forhåndsdefinert rammeverk eller utvalg av kvalitetskriterier (s. 73). Denne metoden har sitt utspring i kvalitetssikring av biblioteks-kataloger, og innebærer at flere eksperter analyserer hvert sitt utvalg, før resultatene sammenstilles for å gi en samlet beregning over hele samlingens metadata-kvalitet. Denne metoden i sin helhet er ikke aktuell for min undersøkelse, men den viser at det kan forsvares å analysere metadata-kvalitet manuelt, gitt at man har faste kriterier å måle kvaliteten ut ifra. I utviklingen av mitt analyseverktøy i det følgende vil disse kriteriene fremkomme.

4.3 Analysemetode/-verktøy

For å analysere kvaliteten på metadata-postene, må jeg velge hvilke metadatafelt det er mest hensiktsmessig å analysere. Utvelgelsen vil gjøres ut ifra FAIR-prinsippene og tidligere forskning referert til i teori-kapittelet, ved å stille følgende spørsmål: Hvilke felt kan imøtekomme kravene i FAIR-prinsippene om tilrettelegging for *gjenfinnbarhet, interoperabilitet* og *gjenbruk?* Hvilke felt viser om kvalitetskriteriene *nøyaktighet, fullstendighet og overensstemmelse* imøtekommes? I hvilke felt legges det til rette for *standardisering/kontrollerte vokabularer*?

4.3.1 Utvelgelse av metadata-felt

Nedenfor vil jeg presentere de metadata-feltene jeg velger å analysere. Dette gjør jeg ved å klargjøre hva slags informasjon feltene kan inneholde, hva som fremkommer i tidligere undersøkelser av det enkelte felt, samt hvordan analysen vil si noe om metadataenes kvalitet målt opp mot FAIR-prinsippene og resultatene av tidligere forskning. Jeg vil også gjøre rede for hva slags type metadata det enkelte felt representerer (deskriptivt, administrativt eller strukturelt), i henhold til Tabell 1 og 2 (Riley, 2017).

Før jeg gå nærmere inn på det enkelte metadata-felt, vil jeg gjøre noen generelle betraktninger.

FAIR-prinsippenes krav om *rikholdige metadata* og litteraturens anbefaling om *fullstendighet* kan jeg måle ved å undersøke om *alle* mine utvalgte felt har innhold. Særlig viktig mener jeg det er å fokusere på felt som i teorien kan standardiseres/ha vokabularer, og som slik imøtekommer prinsippet om presise metadata; som *Creator, Subject/keyword* og *Type*. Grundig utfylling av *Description*-feltet vil også bidra til rikholdige metadata.

Gjennom utvelgelsen av feltene nedenfor, velger jeg også bort noen felt. Det kunne ha vært interessant å undersøke dato-feltene, slik Rousidis et al. (2014) gjorde. Han fant en rekke uoverensstemmelser knyttet til dato-format. I DataverseNO har imidlertid dato et fast format, og feilmarginene er derfor ikke-eksisterende. Feltet for *Contributor* vurderer jeg også å analysere. Dette feltet inneholder de samme navnene som feltet *Creator,* men uten identifikatoren ORCID. Jeg bestemmer meg imidlertid for at det er tilstrekkelig å analysere ett av feltene som inneholder personnavn, og velger da *Creator*-feltet, i og med at jeg da kan undersøke i hvilken utstrekning ORCID er brukt.

I det følgende vil jeg presentere de metadata-feltene jeg velger å analysere:

Identifier

Feltet *Identifier* skal inneholde en unik streng som identifiserer og gir en lenke til ressursen (DataCite Metadata Working Group, 2019). Bruk av globalt unike og bestandige identifikatorer tilfredsstiller kravet i FAIR-prinsippene som legger til rette for *gjenfinnbarhet (F)* (Wilkinson et al., 2016). At identifikatoren er globalt unik innebærer at ingen andre kan

gjenbruke den uten å henvise til de opprinnelige dataene den var knyttet til. At den er bestandig betyr at lenken ikke slutter å virke over tid.

Den mest brukte identifikatoren for digitale ressurser er DOI[15] (Digital Object Identifier). Dette er en identifikator som skiller seg fra en URL, fordi den ikke er knyttet til plasseringen av den digitale ressursen, men identifiserer ressursen direkte. Ifølge GO FAIR (u.å.) har man kommet langt i å publisere FAIR data hvis den unike identifikatoren brukes, fordi det er vanskelig å oppnå de andre FAIR-prinsippene uten en unik identifikator. Farnel og Shiri (2014) konstaterer at DOI er nøkkelen til både gjenfinning, bevaring og sitering av forskningsdata (s. 79). For forskerens egne interesser er det særlig viktig å bli sitert riktig når ens data blir gjenbrukt (GO FAIR, u.å.).

Nødvendigheten av at den elektroniske ressursen har en unik identifikator, og da fortrinnsvis standarden DOI, samsvarer med anbefalingen i tidligere forskning om *standardisering* (Rousidis et al., 2014).

Som metadata-type (Riley, 2017) tilhører *Identifier*-feltet både *deskriptive* og *administrative* metadata. Ved at DOI-nr linker til beskrivelsen av ressursen, er det deskriptivt, mens det er administrativt når det er et redskap for anskaffelse av ressursen (Caplan, 2003, s. 4).

I DataverseNO, som mine data er hentet fra, krever metadata-skjema som brukes at feltet *Identifier* bestå av den unike identifikatoren DOI. Jeg vil derfor undersøke om feltet *Identifier* er utfylt med en DOI.

Creator

Feltet *Creator* inneholder navnene på den eller de som har skapt datasettet. Et viktig aspekt ved lagring og gjenfinning av personnavn, er at navnet har en standardisert form (Kim et al., 2019; Rousidis et al., 2014). I en undersøkelse av studier av ulike forskningsdata-arkiver og deres krav til metadata, avdekket Kim et al. (2019) at ufullstendig utfylling av *Creator*-feltet var et funn som gikk igjen (s. 847). Disse studiene anbefalte derfor å lagre personnavn ved hjelp av en unik forfatter-ID. Rousidis et al. (2014) analyserte blant annet feltet *Creator* i en studie av ulike forskningsdata-arkiver (s. 282). De fant store avvik i form av ulike versjoner

[15] https://www.doi.org/

av navn, som for eksempel varierende bruk av initialer kontra fullstendig navn, manglende fornavn og varianter av navn på grunn av ulike transkriberinger. Konklusjonen ble også her at en unik forfatter-ID/bruk av autoritetsregister er påkrevd, og de nevnte ORCID som et eksempel på en slik ID (Rousidis et al., 2014, s. 285). ORCID er den navnestandarden som anbefales brukt i metadata-skjemaet til DataverseNO.

Alle de tre FAIR-prinsippene som er bakgrunnen for forskningsspørsmålene i denne oppgaven, blir berørt av feltet *Creator* (Wilkinson et al., 2016):

Gjenfinnbarheten (F) styrkes ved at personnavn lagres ved hjelp av en unik og varig identifikator. Dette tilsvarer kravet om *standardisering* i litteraturen. Videre legger *entydige metadata* til rette for gjenfinnbarhet, noe som også kan bety standardisering av personnavn. Dette tilsvarer også kravet om *nøyaktighet* i litteraturen.

Interoperabiliteten (I) styrkes ved bruk av et standardisert språk/vokabular, noe som igjen imøtekommes ved bruk av standardisert navneform. Dette samsvarer med kravet i litteraturen om *overensstemmelse*, det vil si at formen på personnavn er identisk i alle de metadata-postene de forekommer i. Kravet om *nøyaktighet* er også samsvarende med et standardisert språk/vokabular. Videre kommer det klare anbefalinger i litteraturen om å bruke kontrollerte vokabularer.

Muligheten for *gjenbruk (R)* styrkes ved at metadata inneholder presise beskrivelser, noe en standardisert navneform også er et uttrykk for. Dette samsvarer igjen med litteraturens krav om standardisering og nøyaktighet.

Som metadatatype (Riley, 2017) tilhører *Creator*-feltet typen deskriptive metadata. Ifølge Tabell 2 er hensikten med dette feltet "discovery", "display" og "interoperability", hvilket samsvarer med FAIR-prinsippene om gjenfinnbarhet og interoperabilitet som nevnt ovenfor.

Oppsummert vil jeg undersøke utfyllingen av feltet *Creator* for å finne ut i hvor stor grad ORCID er brukt som en autorisert navneform, og for å kartlegge kvaliteten på personnavn som ikke er registrert ved hjelp av ORCID. Dette vil avdekke i hvilken grad metadataene legger til rette for gjenfinning (F); ved bruk av en unik og varig identifikator og entydige

metadata, samt imøtekommelse av kravene i litteraturen om standardisering og nøyaktighet. Videre vil analysen av dette feltet avdekke i hvilken grad metadataene legger til rette for interoperabilitet (I); ved at det brukes et standardisert språk/vokabular, og ved at kravene i litteraturen om overensstemmelse, nøyaktighet og et kontrollert vokabular imøtekommes. Analysen av dette feltet vil også avdekke i hvilken grad metadataene legger til rette for gjenbruk (R), ved at de inneholder en presis beskrivelse, noe som igjen samsvarer med kravene i litteraturen om standardisering og nøyaktighet.

Description

Dette feltet blir brukt til å gi et sammendrag og en beskrivelse av forskningsprosjektet forskningsdataene er grunnlag for, samt informasjon om forskningsdataenes hensikt, egenskaper og omfang. Feltet bør dessuten inneholde vesentlig informasjon om innsamlingsmetode, metode for dataprosessering og analyse, samt kontekst (Choudhury et al., 2018, s. 7). Lengden og detaljnivået på disse tekstene varierer sterkt, og årsaken til dette kan være type prosjekt, type data, disiplin m.m. Kim et al. (2019) påpeker at kravet til bredde og dybde som kreves og forventes i beskrivelsen av forskningsdata varierer mellom disipliner (s. 863). Det vil kreve stor grad av kunnskap innenfor de ulike disiplinene for å kunne vurdere kvaliteten på tekstene i dette feltet. Jeg har derfor valgt å undersøke dette feltet binært, dvs. om det er utfylt eller ikke. Etter min vurdering vil undersøkelsen allikevel avdekke en egenskap ved metadataene, ved å slå fast om feltet har innhold eller ikke.

Analysen av dette feltet berører to av FAIR-prinsippene. *Gjenfinnbarhet (F)* forutsetter rikholdige metadata. Dette korresponderer med tidligere forsknings krav om *fullstendighet* (Kim et al., 2019; Marc et al., 2016; Rousidis et al., 2014). Videre forutsetter tilrettelegging for *gjenbruk (R)* en *grundig beskrivelse,* som også samsvarer med kravet om *fullstendighet.*

Som navnet tilsier, tilhører feltet *Description* metadatatypen deskriptive metadata. I følge Riley (2017) gir deskriptive metadata informasjon om en ressurs for at den skal kunne gjenfinnes og forstås (s. [10]), og innholdet i dette feltet er absolutt ment å skulle bidra til forståelsen av ressursen.

Oppsummert vil jeg undersøke utfyllingen av feltet *Description* for å avdekke om metadata er rikholdige og inneholder en grundig beskrivelse, i henhold til FAIR-prinsippene om

gjenfinning og gjenbruk. Dette samsvarer med tidligere forsknings krav om fullstendighet. Forbeholdet om graden av utbytte ved å analysere dette feltet, er at jeg ikke har mulighet for å analysere dets innhold. Undersøkelsen vil kun avdekke om feltet har innhold eller ikke.

Subject/Keyword

Subject/Keyword-feltet er det en selvfølge å analysere; både fordi det har vært gjenstand for undersøkelse i alle kildene mine, og fordi det er et felt hvor bruk/ikke bruk av kontrollert vokabular har stor betydning, spesielt for interoperabiliteten (Wilkinson et al., 2016). Rousidis et al. (2015) og Balatsoukas et al. (2018) gjorde en studie av *Subject*-feltet i Dublin Cores metadataskjema i forskningsdata-arkivet Dryad for å avdekke eventuelle kvalitetsproblemer. Funnene viste problemer knyttet til manglende kontrollert vokabular og standardisering; for eksempel tilfeldig bruk av entall/flertall, synonymer, ulike ordklasser m.m. (Balatsoukas et al., 2018, s. 1). De pekte også på et viktig aspekt som er unikt for metadata for forskningsdata. Det kan nemlig oppstå flere versjoner av et datasett, og da er det av stor viktighet at emneordene kommer fra kontrollerte vokabularer, taksonomier eller tesauri, slik at den semantiske sammenhengen mellom ulike datasett og innenfor et datasetts livssyklus blir ivaretatt (Balatsoukas et al., 2018, s. 7).

Analysen av dette feltet berører alle de tre FAIR-prinsippene *gjenfinnbarhet (F), interoperabilitet (I) og gjenbruk (R).* Dette innebærer henholdsvis kravet om *entydige metadata* (F), i samsvar med kravet om *nøyaktighet* i tidligere forskning, kravet om *standardisert språk/vokabularer* (I), i samsvar med kravet om *overensstemmelse, nøyaktighet* og *kontrollerte vokabularer* i tidligere forskning, og kravet om en *presis beskrivelse* (R), i samsvar med litteraturens krav om *standardisering* og *nøyaktighet*.

Emne-feltene hører til metadata-typen deskriptive metadata, da de bidrar til å forstå innholdet i ressursen og kan legge til rette for gjenfinning og interoperabilitet (Riley, 2017).

Da *Subject*- og *Keyword*-feltene har ulike egenskaper i mine data, vil jeg analysere dem separat. I mitt datasett i formatet Dublin Core heter alle emneordsfeltene *Subject*, mens det i visningsformatet på søkesiden til DataverseNO er brukt begge benevnelsene. Fra Universitetet i Bergen har jeg fått vite at ved registrering av metadata i DataverseNO har *Subject* faste verdier, mens *Keyword* er et fritekstfelt (B.K. Humberset, personlig kommunikasjon, 22. jan.

2020). Ved å holde markøren over spørsmålstegnet ved siden av navnet på det enkelte felt i en tilfeldig valgt metadatapost hos DataverseNO, vises beskrivelsen av feltene *Subject* og *Keyword* som henholdsvis "Domain-specific Subject Categories that are topically relevant to the Dataset" og "Key terms that describe important aspects of the Dataset".

Oppsummert vil jeg analysere feltene Subject og Keyword for å undersøke hvorvidt metadataene er entydige, standardiserte og presise i sin beskrivelse, det vil si om feltene er utfylt ved hjelp av kontrollerte vokabularer. Dette samsvarer med forskningens krav om nøyaktighet, overensstemmelse, standardisering og bruk av kontrollerte vokabularer.

Type

Feltet *Type*, eller *Kind of data* som det kalles på DataverseNO, brukes til å angi hva slags type data som finnes i de lagrede filene med forskningsdata. Feltet er fritekst (B.K. Humberset, personlig kommunikasjon, 11. febr. 2020). Ved å holde markøren over spørsmålstegnet knyttet til feltets navn i en tilfeldig valgt metadatapost hos DataverseNO, får man opp en rekke eksempler på hvilke verdier dette feltet kan ha: *survey data, sensus/enumeration data, aggregated data, clinical data, event/transaction data, program source code, machine-readable text, administrative records data, experimental data, psychological test, textual data, coded textual, coded documents, time budget diaries, observation data, ratings, process-produced data, or other.*

Dette feltet er undersøkt i to av kildene mine (Kim et al., 2019; Rousidis et al., 2014), og er et felt som er spesielt viktig for forskningsdata. Å få informasjon om hva slags type forskningsdata metadataene beskriver, er vesentlig i forhold til muligheten for gjenbruk. Dessuten avdekker tidligere forskning at inkonsekvens/manglende standardisering også her er en utfordring (Kim et al., 2019). Rousidis et al. (2014) analyserte utfyllingen av blant annet feltet *(Resource)type* i The Dryad data repository, og fant både tomme felt og felt med irrelevante verdier. De anbefalte en liste med faste verdier å velge mellom for å garantere for kvalitet i utfyllingen av *Type*-feltet (s. 284-285).

Analysen av dette feltet berører de samme FAIR-prinsippene som feltene for *Subject* og *Keyword*, nemlig *gjenfinning (F), interoperabilitet (I)* og *gjenbruk (R)*. *Gjenfinning (F)*

forutsetter *entydige* metadata, noe som eventuelle faste verdier i *Type*-feltet vil legge til rette for. Dette samsvarer med tidligere forsknings krav om *nøyaktighet*. Videre vil eventuelt standardisert språk/vokabularer legge til rette for *interoperabilitet (I)*, noe som korresponderer med litteraturens krav om *overensstemmelse, nøyaktighet og kontrollerte vokabularer*. *Gjenbruk (R)* forutsetter en presis beskrivelse, uttrykt i litteraturen som *standardisering* og *nøyaktighet*. Som nevnt er det spesielt med tanke på gjenbruk viktig for forskeren å vite hvilken datatype han har med å gjøre.

Dette feltet tilhører også metadata-typen deskriptive metadata, da det gir informasjon om innholdet i ressursen og kan legge til rette for gjenfinning og interoperabilitet (Riley, 2017).

Oppsummert er det primære målet med å undersøke dette feltet å avdekke om det er utfylt ved hjelp av faste verdier, for slik å legge til rette for entydighet, standardisering og en presis beskrivelse. Dernest blir fokus å analysere kvaliteten på begrepene som er brukt.

Lisence

Lisence er et felt som er spesielt for forskningsdata. Det er av stor betydning, da det gir informasjon om hvorvidt forskningsdataene er fritt tilgjengelige for gjenbruk (Choudhury et al., 2018). Hensikten med å analysere dette feltet er å kartlegge hva slags lisens, dvs. mulighet for gjenbruk, postene er utstyrt med. Creative Commons er en internasjonal, ideell organisasjon som utvikler og stiller til rådighet et sett med kvalitetssikrede lisenser for skapere av åndsverk, produsenter og utøvere[16]. De har utviklet verktøyet Creative Commons Zero (CC0)[17], som indikerer at man gir fri tilgang til gjenbruk av ens data, også utenfor Norge.

Analysen av dette feltet berører FAIR-prinsippet *gjenbruk (R)*, som fremsetter kravet om en fri og tilgjengelig lisens for bruk (Wilkinson et al., 2016). Ifølge RDA-Codata Legal Interoperability Interest Group (2016) blir også den juridiske interoperabiliteten, nevnt i avsnittet om interoperabilitet tidligere i denne oppgaven, styrket ved at metadataposten har verdien CC0 i feltet for *Lisence*.

[16] https://no.wikipedia.org/wiki/Creative_Commons
[17] https://creativecommons.org/share-your-work/public-domain/cc0/

Dette feltet tilhører metadata-typen administrative metadata, som igjen er inndelt i tre undertyper; technical-, preservation- og rights metadata (Riley, 2017). *Lisence* faller inn under typen rights metadata, og i Tabell 2 er *Lisence terms* et eksempel på en av egenskapene dette feltet kan ha.

Oppsummert vil jeg undersøke om feltet *Lisence* har verdien CC0, for slik å legge til rette for gjenbruk (R).

4.3.2 Analyseverktøy

Ved hjelp av definisjoner og modeller av begrepene i oppgavens problemstilling, sammen med tidligere forskning og deler av FAIR-prinsippene, har jeg vist hvordan jeg har valgt ut de metadata-feltene det er mest hensiktsmessig å analysere, samt hvilke kriterier jeg skal bruke i analysen av dem. På denne bakgrunnen har jeg bygget ut et analyseverktøy som jeg kan bruke på oppgavens datasett.

For å gi et overblikk over analysen jeg skal foreta, har jeg laget følgende tabell (Tabell 5), der første kolonne består av de tre FAIR-prinsippene som angår kvaliteten på metadata. Dette tilsvarer mine tre forskningsspørsmål. Andre kolonne beskriver kravene til metadata som hver av disse prinsippene fremmer, slik de er blitt gjennomgått ovenfor. Tredje kolonne består av kvalitetskriterier fremkommet i tidligere forskning, plassert i overensstemmelse med det enkelte krav i FAIR-prinsippene. I den fjerde kolonnen har jeg plassert de metadata-feltene som skal undersøkes med tanke på det enkelte kvalitetskrav.

FAIR	FAIRs metadata-krav	Tidligere forskning	Metadata-felt
F1 **Findable**	Unik og varig ID	Standardisering	Identifier Creator
	Entydige metadata	Standardisering Nøyaktighet	Creator Subject Keyword Type
	Rikholdige metadata	Fullstendighet	Description
F2 **Interoperable**	Standardisert språk/ vokabularer	Overensstemmelse Nøyaktighet Kontrollerte vokabularer	Creator Subject Keyword Type
F3 **Reusable**	Grundig beskrivelse	Fullstendighet	Description
	Presis beskrivelse	Standardisering Nøyaktighet	Creator Subject Keyword Type
	Klar og tilgjengelig lisens for bruk		Lisence

Tabell 5: Analyseverktøy ordnet etter FAIR-prinsippene.

Ut ifra tabellen ovenfor vet jeg hvilke kvaliteter jeg skal se etter når jeg undersøker utfyllingen av det enkelte felt. På det grunnlaget har jeg laget en ny tabell (Tabell 6) ordnet etter metadatafelt, og med tilhørende spørsmål som skal undersøkes i kolonne 2. I den tredje kolonnen har jeg plassert de FAIR-prinsippene som skal imøtekommes, med utdyping av disse kravene i kolonne 4. Til sist kommer en kolonne med kvalitetskravene fra tidligere forskning. Ved hjelp av denne tabellen kan jeg analysere datasettet og kartlegge kvaliteten på metadataene.

Metadatafelt	Undersøke	FAIR	FAIRs metadata-krav	Tidligere forskning
Identifier	DOI?	F	Unik og varig ID	Standardisering
Creator	ORCID?	F	Unik og varig ID Entydige metadata	Standardisering Nøyaktighet
		I	Standardisert språk/ vokabularer	Overensstemmelse Nøyaktighet Kontrollerte vokabularer
		R	Presis beskrivelse	Standardisering Nøyaktighet
Description	Er feltet utfylt?	F	Rikholdige metadata	Fullstendighet
		R	Grundig beskrivelse	Fullstendighet
Subject	Kontrollert vokabular?	F	Entydige metadata	Nøyaktighet
		I	Standardisert språk/ vokabularer	Overensstemmelse Nøyaktighet Kontrollerte vokabularer
		R	Presis beskrivelse	Standardisering Nøyaktighet
Keyword	Kontrollert vokabular?	F	Entydige metadata	Nøyaktighet
		I	Standardisert språk/ vokabularer	Overensstemmelse Nøyaktighet Kontrollerte vokabularer
		R	Presis beskrivelse	Standardisering Nøyaktighet
Type	Faste verdier?	F	Entydige metadata	Nøyaktighet
		I	Standardisert språk/ vokabularer	Overensstemmelse Nøyaktighet Kontrollerte vokabularer
		R	Presis beskrivelse	Standardisering Nøyaktighet
Lisence	CC0?	R	Klar og tilgjengelig lisens for bruk	

Tabell 6: Analyseverktøy ordnet etter metadatafelt som skal undersøkes.

5 Resultater

Datasettet mitt ble hentet fra DataverseNO og består av 131 metadata-poster som ble publisert i tidsrommet 01.01.2018 – 23.03.2020. Postene er publisert av åtte ulike institusjoner, og fordeler seg som følger i antall: INN: 4, NMBU: 8, NTNU: 7, TGO: 1, TROLLing: 24, UIA: 5, UIB: 5 og UIT: 77. Årsaken til at overveiende mange av postene er publisert av UIT, er at UIT er eiere av DataverseNO og derfor antagelig bruker dette arkivet mer enn de andre institusjonene som først har begynt å bruke DataverseNO i senere tid. TROLLing tilhører også UIT, og har en egen samling innen DataverseNO, noe som forklarer hvorfor de også er overrepresentert i datasettet.

Det må bemerkes at jeg, som tidligere nevnt, har fjernet en rekke nesten identiske poster der årstallet i tittelen var den eneste ulikheten. Dette fordi hver og en av dem ikke hadde noe unikt å tilføre undersøkelsen, samtidig som de ville skape en skjev fordeling i funnene. Fra det opprinnelige datasettet bestående av 590 poster, fjernet jeg følgende poster: 40 av 41 poster publisert av UIT med tittelen "Hydrographic data from Northern Norwegian fjords – […]", 17 av 18 poster publisert av UIT med tittelen "Photo frames on non-permanent stations […]", 29 av 30 poster publisert av UIT med tittelen "Scenery photos without the frame […]" og 373 av 374 poster publisert av TGO med tittelen " TGO Ramfjordmoen Ionosonde Data […]". Til sammen utgjør dette 459 fjernede poster, og et resterende datasett bestående av 131 poster. Sortert etter publiserings-år blir fordelingen som følger: 2018: 39 poster, 2019: 65 poster og 2020: 27 poster.

Metadata-postene er høstet fra DataverseNO i formatet Dublin Core. Under (Figur 5) vises et eksempel på en av postene, der de feltene som skal analyseres er merket med gult. Dette er de feltene som ble lastet opp i en CSV-fil og lagt inn i et Excel-dokument.

Det viser seg at det ikke er mulig å finne ut om forfatter er registrert med ORCID ut ifra disse dataene. I tillegg må jeg derfor søke opp hver enkelt post hos DataverseNO for å finne denne informasjonen. I Figur 6 vises den samme posten som i Figur 5, men nå slik den fremstår ved søk. De aktuelle feltene, samt informasjonen om ORCID, er merket med gult. Personnavn i

feltene *Creator* og *Contributor* er skjult på grunn av personvernhensyn, etter instrukser fra Norsk senter for forskningsdata (NSD)[18].

Figur 5: Eksempel på metadata-post fra datasettet, i formatet Dublin Core.

[18] https://nsd.no/personvernombud/

Figur 6: Eksempel på metadata-post fra datasettet, hentet fra https://dataverse.no/.

I det følgende vil jeg presentere mine funn, ut ifra spørsmålene jeg har stilt til hvert metadatafelt i analyseverktøyet (Tabell 6). Funnene presenteres i den rekkefølgen feltene kommer i metadatapostene, for så å oppsummeres i Tabell 7. Deretter blir resultatene analysert i henhold til oppgavens forskningsspørsmål og problemstilling.

5.1 Identifier

I feltet *Identifier* er det forventet å finne en DOI (Digital Object Identifier), og samtlige poster i datasettet har en DOI, det vil si en globalt unik og varig identifikator.

5.2 Creator

I DataverseNO generelt er forskerens ORCID brukt i stor utstrekning, og blir anbefalt i retningslinjene deres (DataverseNO, u.å.). Dette er en alfanumerisk kode som gir personer en unik identifikator med en autorisert navneform. I 115 av i alt 131 poster i datasettet er forfatterne personer. Det er dermed i disse postene det er mulig å finne bruk av ORCID. I 64 poster er samtlige navn registrert med ORCID. Av de resterende 51 postene er samtlige navn uten ORCID i 35 poster, mens noen navn er med og noen navn uten ORCID i de andre 16 postene.

Til sammen er 112 av totalt 282 navn, det vil si nærmere 40 %, registrert uten ORCID. Jeg søker opp disse navnene på www.orcid.com for å undersøke om vedkommende er registrert der. Hele 96 navn, det vil si nærmere 86 % av de 112 navnene uten ORCID, har en ORCID som ikke er tatt i bruk i postene.

Ved å sortere alle forfatternavn alfabetisk, avdekker jeg bruk av ulike navneformer. I noen tilfeller er initialer brukt i stedet for mellomnavn, mens mellomnavnet på samme person er skrevet helt ut i andre tilfeller. Totalt finner jeg seks slike tilfeller. Andre svakheter er ni navn som ikke er skrevet i invertert form. Det mest oppsiktsvekkende funnet er imidlertid at også navn registrert med ORCID i noen tilfeller har en annen form enn formen registrert hos www.orcid.com. Dette tyder på at registrering av personnavn og registrering av ORCID-nr er to separate operasjoner; det vil si at man ikke har hentet inn den autoriserte ORCID-navneformen, men kun har lagt inn navnet slik man selv har ønsket å skrive det.

5.3 Description

Målet med å analysere dette feltet, er å undersøke om det har innhold, og samtlige poster i datasettet har innhold i dette feltet. Som tidligere nevnt er det ikke mulig å analysere tekstene i dette feltet, da det krever omfattende fagkunnskap innenfor den enkelte disiplin.

5.4 Subject/keyword

Subject-feltet i datasettet er konsekvent fylt ut med faste verdier.

Keyword-feltet er fritekst. Fordelt på de 131 postene finnes det 606 unike emneord. Emneordene bærer preg av at de er fritekst og skapt av den enkelte forsker; med ulike former av samme ord (for eksempel "seismic" og "seismics"), mer eller mindre spesifisitet av samme emne/fenomen (for eksempel "water" og "water masses"), synonymer, for generelle begreper (for eksempel "development"), flere emneord i samme felt, samt bruk av de faste verdiene fra *Subject*-feltet.

5.5 Type

Av de 131 postene i datasettet, har 54 poster (41 %) ikke oppgitt type. 16 av de 77 resterende postene har brukt noen av de foreslåtte verdiene fra DataverseNO. De andre 66 postene har brukt 51 unike verdier. En del av disse termene handler mer om innholdet i dataene enn data-type, og burde ha vært plassert i *Keyword*-feltet i stedet; for eksempel "methane concentration", "methane free gas flow rate", "pressure/depth", "vegetation cover". Andre svakheter er ulike former av samme begrep, for eksempel "simulated data" og "simulation data"; "survey data" og "survey-based data".

5.6 Lisence

Alle postene i datasettet har koden CC0 i feltet for *Lisence*, og gir således fri tilgang til gjenbruk.

5.7 Oppsummering

For å gi en bedre oversikt over resultatene, setter jeg nedenfor en oppsummert versjon av resultatene inn i en tabell (Tabell 7). I venstre kolonne angis de metadata-feltene jeg har undersøkt. Neste kolonne angir i korthet hva jeg skulle undersøke. Tredje kolonne angir funnene binært, der det var mulig, mens siste kolonne gir en meget forkortet beskrivelse av funnene.

Metadatafelt	Undersøke	Binære funn	Beskrivelse av funn
Identifier	Er feltet utfylt?	Ja	Alle postene har en DOI.
Creator	Unik ID?		Delvis bruk av ORCID, og en del varierende navneformer.
Description	Er feltet utfylt?	Ja	Alle postene hadde en kortere eller lengre tekst i dette feltet.
Subject	Kontrollert vokabular?	Ja	Faste verdier å velge mellom.
Keyword	Kontrollert vokabular?	Nei	Fritekst-felt Synonymer, ulike former av samme ord m.m.
Type	Faste verdier?	Nei	Fritekst-felt 41 % uten verdi Ulike former av samme begrep Noen verdier som egentlig er emneord
Lisence	CC0?	Ja	Alle postene har en CC0.

Tabell 7: Resultater

6 Analyse og tolkning

I det følgende vil jeg analysere undersøkelsens resultater opp mot analyseverktøyet (Tabell 5 og 6), samtidig som jeg tolker resultatene ved å svare på oppgavens forskningsspørsmål. Deretter vil jeg sammenfatte resultatene fra analyse og tolkning ved å svare på oppgavens problemstilling: *Hva kjennetegner kvaliteten på norske metadata for forskningsdata?*

6.1 I hvilken grad fremmer metadataene gjenfinnbarhet?

Ifølge oppgavens analyseverktøy, måles metadataenes grad av gjenfinnbarhet ved å undersøke om de er tildelt en *unik og varig ID*, noe som tilsvarer kravet om standardisering i tidligere forskning; videre om metadataene er *entydige*, som korresponderer med kravene om standardisering og nøyaktighet i tidligere forskning; samt om metadataene er *rikholdige*, tilsvarende kravet om fullstendighet i tidligere forskning. I det følgende vil jeg analysere i hvilken grad undersøkelsens resultater imøtekommer disse kravene, og konkludere med hvilken innflytelse dette har på gjenfinnbarheten.

6.1.1 Unik og varig ID

Det første kriteriet som må være oppfylt for at forskningsdata skal være gjenfinnbare, er at data og metadata er tildelt en *globalt unik og bestandig identifikator* (Wilkinson et al., 2016). Alle metadata-postene i datasettet har en DOI (Digital Object Identifier), og legger således til rette for at dataene er gjenfinnbare. Dette er, som tidligere nevnt, det viktigste prinsippet, da det er umulig å innfri de andre FAIR-prinsippene uten unike og vedvarende identifikatorer (GO FAIR, u.å.). DOI er vesentlig for sitering og referanser, samt koblingen mellom datasett og metadata (Hodson et al., 2018, s. 19), og ifølge GO FAIR (u.å.) er man ved å bruke en slik identifikator allerede kommet langt på vei i å publisere FAIR data. Farnel og Shiri (2014) beskriver bruken av DOI som "key to discovery, preservation and citation of research data" (s. 79). Tidligere forsknings krav om standardisering blir definitivt imøtekommet ved bruk av standarden DOI.

Identifikatoren ORCID for navn på *creator* er kun delvis tatt i bruk i mine data, i mange tilfeller ved registrering av ORCID-nr uten den tilhørende standardiserte navneformen. Dette svekker målet om å bruke en unik og varig ID og legger ikke til rette for standardisering. De

samme svakhetene ble avdekket i undersøkelsene som Kim et al. (2019) og Rousidis et al. (2014) utførte.

6.1.2 Entydige metadata

At forskningsdataene er lagret ved hjelp av *entydige* metadata er det andre kriteriet for gjenfinnbarhet (Wilkinson et al., 2016). *Entydighet* har med standardisering og nøyaktighet å gjøre, og i min undersøkelse er det feltene *Creator, Subject, Keyword* og *Type* som kan standardiseres. Bortsett fra feltet *Subject,* som har faste verdier, har de andre tre feltene svakheter på dette området fordi de er fritekst-felter.

Graden av gjenfinning kan svekkes ved at poster ikke blir funnet fordi forfatternavnet i *Creator*-feltet er feilskrevet/ufullstendig. Funnene som avdekket ufullstendig utfylling av *Creator*-feltet i undersøkelsene til Kim et al. (2019) og Rousidis et al. (2014), samsvarer med mine funn, det vil si en rekke tilfeller av blant annet ufullstendige navneformer og tilfeldig bruk av initialer, kontra navn som var skrevet fullt ut. De anbefalte bruk av en unik forfatter-ID, for eksempel ORCID, for å sørge for standardisering av navneform. Min undersøkelse viser imidlertid at bruk av ORCID ikke har ført til standardiserte navneformer, da personnavn ser ut til å være registrert uavhengig av ORCID-formen.

Fraværet av standardisering/faste vokabularer i *Keyword*-feltet svekker også gjenfinnbarheten. Litteraturstudien til Rousidis et al. (2015) viste at emnefeltet var et av de mest utfordrende områdene både for metadata-skaping og gjenfinning. Hovedårsaken til dette var uerfarne metadata-skapere som ikke hadde forutsetninger for å velge gode og uproblematiske emneord, og resultater fra flere case-studier viste at for å oppnå høy kvalitet på emneordene burde forfatter og metadata-spesialist samarbeide (Rousidis et al., 2015, s. 206). I sin undersøkelse av ulike forskningsdata-arkiver konkluderer dessuten Kim et al. (2019) med at metadata som skal registreres ved deponering av forskningsdata burde defineres på en mer standardisert måte – spesielt de som kan ha ulike betydninger innen og mellom disipliner (s. 865). Å søke opp disse postene ved hjelp av emneord kan bli en utfordring, spesielt hvis man søker innen et fagfelt som ikke er ens eget. Da hadde det vært en stor fordel å ha et fast vokabular å ta utgangspunkt i.

Feltet *Type* er heller ikke utfylt på en *entydig* måte, og svekker således muligheten for gjenfinning. Standardisering i form av faste verdier ville ha styrket entydigheten betraktelig, noe Rousidis et al. (2014) også konkluderer med når de hevder at "in the case of the Type metadata element, insistencies can be fixed through the use of pre-defined lists of values for authors to select from" (s. 285).

DataCite[19] er en annen tjeneste som tilgjengeliggjør forskningsdata fra mange ulike arkiver. For å få et sammenligningsgrunnlag undersøker jeg deres skjema, og i denne sammenheng retningslinjene for utfylling av *Type*-feltet (DataCite Metadata Working Group, 2019). De anbefaler en *resourceTypeGeneral value* med kontrollerte verdier, og i konteksten forskningsdata er "Dataset" den aktuelle verdien å bruke. Dernest anbefaler de å tilføye en sub-property, en såkalt *resourceType value*, som spesifiserer hva slags type data som er lagret. Denne termen kan være fritekst eller hentet fra en tesaurus.

Med en liste med faste entydige verdier å velge mellom, ville søk på datatype og følgelig gjenfinning blitt betraktelig styrket.

6.1.3 Rikholdige metadata

At metadata er *rikholdige* er nærmere utdypet som at de er beskrevet med et mangfold av presise og relevante egenskaper (Wilkinson et al., 2016), tilsvarende tidligere forsknings krav om fullstendighet. Generelt innebærer dette at alle feltene i metadataskjema er fylt ut, noe som bare delvis er tilfellet i min undersøkelse. Feltet *Description* har innhold i samtlige poster, og dette er en styrke for gjenfinnbarheten, da det er et søkbart felt som kan gi detaljert informasjon om datasettet. At jeg ikke har kunnet analysere innholdet i dette feltet begrenser imidlertid min mulighet for å konkludere hvorvidt metadataene er rikholdige eller ikke. Feltet for *Type* har bare innhold i 77 av 131 poster, og svekker således graden av rikholdige metadata ved at 41 % av postene ikke er gjenfinnbare ved søk på datatype.

6.1.4 Konklusjon

Oppsummert blir gjenfinnbarheten styrket ved gjennomført bruk av DOI, noe som er den viktigste egenskapen ved metadataene. Standardisering av *Subject*-feltet gir også økt gjenfinning, ved at man kan søke på standardiserte benevnelser på de ulike fagfeltene. Utover

[19] https://datacite.org/

disse funnene er mangelen på standardisering av *Keyword, Type* og delvis *Creator* de store svakhetene som svekker gjenfinnbarheten. Svaret på F1 blir derfor at metadataene i denne undersøkelsen bare delvis fremmer gjenfinnbarhet.

6.2 I hvilken grad fremmer metadataene interoperabilitet?

Ifølge oppgavens analyseredskap måles metadataenes grad av interoperabilitet ved å undersøke om de bruker *et standardisert språk/vokabularer* (Wilkinson et al., 2016). Dette tilsvarer kravene om overensstemmelse, nøyaktighet og kontrollerte vokabularer i tidligere forskning. I det følgende vil jeg analysere i hvilken grad undersøkelsens resultater imøtekommer disse kravene, og konkludere med hvilken innflytelse dette har på interoperabiliteten.

6.2.1 Standardisert språk/vokabularer

Fire av feltene som berører grad av gjenfinnbarhet (*Creator, Subject, Keyword* og *Type*) påvirker også graden av interoperabilitet. Interoperabilitet handler om at metadata skal kunne krysse grenser mellom ulike informasjonskontekster (Haslhofer & Klas, 2010, s. 13). For at ulike systemer skal kunne forstå, lagre og analysere metadataene, er standardisering meget viktig. Kim et al. (2019) konstaterte at interoperabiliteten blir svekket når emneordene ikke er hentet fra et kontrollert vokabular. Følgelig legger de faste verdiene i *Subject*-feltet til rette for interoperabilitet, mens det mye mer brukte *Keyword*-feltet, som er et fritekst-felt, hindrer interoperabiliteten. Fritekst-feltet *Type* hemmer også interoperabiliteten, samt *Creator*-feltet i de tilfellene der det finnes uoverensstemmelser mellom navneformer.

6.2.2 Konklusjon

Ut ifra svakhetene avdekket i denne undersøkelsen, må det konkluderes med at interoperabiliteten, som i stor grad er avhengig av standardisering og vokabularer, ikke blir fremmet i tilstrekkelig grad.

6.3 I hvilken grad fremmer metadataene gjenbruk?

Ifølge oppgavens analyseredskap måles metadataenes tilretteleggelse for gjenbruk ved å undersøke om de gir en *grundig beskrivelse*, noe som tilsvarer kravet om fullstendighet i tidligere forskning; videre om metadataene gir en *presis beskrivelse*, som korresponderer med

kravene om standardisering og nøyaktighet i tidligere forskning; samt om metadataene inneholder en *klar og tilgjengelig lisens for bruk* (Wilkinson et al., 2016). I det følgende vil jeg analysere i hvilken grad undersøkelsens resultater imøtekommer disse kravene, og konkludere med hvilken innflytelse dette har på muligheten for gjenbruk.

6.3.1 Grundig beskrivelse

Grundighet innebærer en fyldig beskrivelse av forskningsdataene i *Description*-feltet. Som nevnt er det ikke mulig å analysere innholdet i dette feltet, men det at det faktisk *har* innhold legger til rette for gjenbruk, ved at forskeren får informasjon om for eksempel dataenes innhold, opphav, kontekst, grad av bearbeidelse osv. (Hodson et al., 2018).

6.3.2 Presis beskrivelse

At metadata er *presise*, betyr et standardisert språk ved bruk av vokabularer. Bruken av faste verdier i *Subject*-feltet i datasettet fremmer dermed gjenbruk, fordi dataene da enklere vil forstås i forskerfellesskapene, som Kunnskapsdepartementet (2017, s. 25) påpeker viktigheten av. Det motsatte er tilfellet med *Keyword*-feltet som er fritekst. Her er det opp til den enkelte forsker å velge emneord som skal beskrive datasettet, og en forståelse på tvers av disipliner begrenses. Slik er det også med feltet *Type,* som også er fritekst og har et tilsynelatende tilfeldig utvalg av verdier. *Type*-feltet er et felt som er spesielt viktig for forskningsdata. Å få informasjon om hva slags type data metadataene beskriver, er vesentlig for mulig gjenbruk. I mine data mangler 41 % av postene innhold i dette feltet. De feltene som har innhold, består av begreper av varierende kvalitet. Mangelfull utfylling av dette feltet begrenser derfor muligheten for gjenbruk.

6.3.3 Klar og tilgjengelig lisens for bruk

Det neste kriteriet for gjenbruk er at forskningsdata har en *klar og tilgjengelig brukerlisens*, slik at de kan brukes fritt uten at man må søke om tillatelse hos data-skaper. Hodson et al. (2018) utdyper at "the conditions under which the data can be used should be transparent to both humans and machines" (s. 20). Alle metadatapostene i datasettet mitt har verdien CC0 i *Lisence*-feltet, og imøtekommer således dette kriteriet. Slik fremmes muligheten for gjenbruk.

6.3.4 Konklusjon

I mine data fremmes gjenbruk gjennom optimal tildeling av brukerlisens, bruk av standardiserte emneord i *Subject*-feltet, samt innhold i *Description*-feltet. På den andre siden er det en stor svakhet med manglende standardisert vokabular for feltene *Keyword* og *Type*. Svaret på dette forskningsspørsmålet blir derfor at metadataene bare delvis fremmer gjenbruk.

6.4 Hva kjennetegner kvaliteten på norske metadata for forskningsdata?

På bakgrunn av svarene jeg har kommet frem til på mine tre forskningsspørsmål, vil jeg vende tilbake til problemstillingen: *Hva kjennetegner kvaliteten på norske metadata for forskningsdata?*

Oppsummert viser svarene på forskningsspørsmålene at mitt utvalg av norske metadata for forskningsdata bare delvis fremmer gjenfinning, interoperabilitet og gjenbruk. Samtlige poster har en DOI som gir en unik og varig identifikator og en CC0-lisens som gir full tilgang til gjenbruk. Dette er viktige standardiserte verdier i metadata for forskningsdata som er avgjørende for henholdsvis gjenfinning og gjenbruk. At *Subject*-feltet legger til rette for faste kategorier er også en styrke som fremmer både gjenfinning, interoperabilitet og gjenbruk. *Description*-feltet har innhold i samtlige poster, og dette styrker muligheten for gjenbruk. Alle disse momentene styrker kvaliteten på metadataene, da de fremmer gjenfinning, interoperabilitet og gjenbruk.

På den andre siden finnes det svakheter som svekker kvaliteten på metadataene, fordi de ikke imøtekommer FAIR-prinsippene og kravene fra tidligere forskning. Mangelen på standardisering/vokabularer er fellesnevneren for disse svakhetene, først og fremst i *Keyword*- og *Type*-feltet. Dette er det samme som kommer frem i tidligere forskning. Kim et al. (2019) avdekket blant annet inkonsekvent utfylling av *Type*-feltet og mangel på standardisert vokabular i *Keyword*-feltet. Rousidis et al. (2015) og Balatsoukas et al. (2018) studerte *Keyword*-feltet og fant kvalitetsproblemer på grunn av manglende standardisert vokabular. Rousidis et al. (2014) foreslo en fast meny å velge termer fra for å avhjelpe problemene med uregelmessigheter i *Type*-feltet. *Creator*-feltet legger også til rette for standardisering i form av ORCID-nr., slik Rousidis et al. (2014, s. 285) anbefalte. ORCID var bare delvis tatt i bruk i mine data, og da på en ufullstendig måte, ved at ORCID-nr ble registrert, men ikke den

tilhørende standardiserte navneformen. Dette er absolutt en svakhet ved kvaliteten på metadataene.

Som beskrevet i teorier om metadata tidligere i oppgaven, deler Haslhofer og Klas (2010) opp metadata i tre "byggesteiner"; (1) *Metadata instance*, som er innholdet i metadata-posten; (2) *Metadata schema*, som er de ulike feltene/elementene i metadata-posten; og (3) *Schema Definition Language*, som er formatet metadata-posten er lagret i (XML, RDF m.fl.). Alle disse tre byggesteinene er med og påvirker kvaliteten på metadataene. Farnel og Shiri (2014) fokuserte på byggestein (2) *Metadata schema* da de sammenlignet bruken av ulike metadata-skjemaer. I min analyse er utgangspunktet å vurdere kvalitet ut ifra byggestein (1) *Metadata instance*, dvs. hvordan feltene i metadata-skjemaet er fylt ut. Det viser seg imidlertid at jeg også avdekker svakheter ved kvaliteten som har sin årsak i skjemaets oppbygning, hvilket betyr at jeg i realiteten også analyserer byggestein (2) *Metadata schema.* Om feltene har faste verdier, bestemmes av skjemaet (byggestein 2). Dette gjelder feltene for *Keyword* og *Type.* At personnavn ikke kan legges inn i *Creator*-feltet med navneformen fra ORCID bestemmes også sannsynligvis av skjemaet. Dermed er det ikke metadata-skaper som har noen påvirkning på hvorvidt standardiserte former blir brukt i disse tre feltene. Utfyllingen av feltene (byggestein 1) påvirkes imidlertid av metadata-skaper, men det ligger utenfor problemstillingen for denne oppgaven å diskutere hvordan dette skal løses i forhold til hvem som skal skape og kuratere metadataene og hvilken kompetanse de skal inneha.

Det finnes med andre ord flere nivåer som påvirker metadata-kvaliteten; metadata-skaper påvirker innholdet i metadata-posten (byggestein 1), men står ikke helt fritt fordi skjemaets utforming legger visse føringer for hva det er mulig å legge inn av innhold (byggestein 2). At det finnes et standardisert metadata-skjema påvirker også kvaliteten på metadataene, og da utelukkende i positiv retning. I sin undersøkelse av ulike forskningsdata-arkiver, konkluderer Austin et al. (2016) med at bruk av ikke-standardiserte metadata-skjemaer hindrer interoperabilitet med andre systemer og ressurser (s. 26). Greenberg Greenberg et al. (2009) påpeker dessuten at Dublin Core, som er det skjemaet DataverseNO bruker, er et enkelt skjema som legger til rette for interoperabilitet ved at det støtter høsting av metadata, søk mellom systemer og utveksling av metadata med andre formater (s. 198). Det at norske metadata for forskningsdata er lagret i dette skjemaet, er dermed et kvalitetsstempel i seg selv.

Det finnes ikke noe entydig svar på denne oppgavens problemstilling. På noen viktige områder kan kvaliteten i det undersøkte datasettet måle seg med FAIR-prinsippene, mens den på andre områder kommer til kort. Det viktigste kvalitetsstempelet ved norske metadata for forskningsdata, er at de alle er tildelt en DOI og en CC0-lisens, og at de er lagret i et standardisert skjema. Den store svakheten er mangel på standardisering/kontrollerte vokabularer, noe som poengteres som en gjennomgående utfordring også i tidligere forskning. Samtidig blir viktigheten av nettopp standardisering også understreket, ikke minst i forhold til mulighetene for interoperabilitet. Kvaliteten på norske metadata for forskningsdata kan derfor karakteriseres som variert, med lignende styrker og svakheter som andre lands metadata for forskningsdata.

7 Konklusjon og diskusjon

Målet med denne undersøkelsen har vært å undersøke kvaliteten på norske metadata for forskningsdata. Det har blitt gjort ved å laste ned et datasett fra det generiske arkivet DataverseNO, for så å analysere metadataene ved hjelp av et analyseverktøy som ble utviklet ved å sammenstille FAIR-prinsippene med tidligere forskning om metadata-kvalitet generelt, samt forskning om kvaliteten på metadata for forskningsdata.

I dette kapittelet vil jeg først se nærmere på de svarene jeg har fått, sett i sammenheng med tidligere forskning, for deretter å evaluere henholdsvis undersøkelsen og analyseverktøyet. Til slutt vil jeg presentere noen perspektiver som kan anspore til videre forskning.

7.1 Konklusjon

Resultatene av undersøkelsen viser høy metadata-kvalitet på enkelte områder, som bruk av DOI og et standardisert metadata-skjema, samt en lisens som gir fri tilgang til gjenbruk, og lavere kvalitet på andre områder, først og fremst på grunn av manglende standardisering/bruk av faste vokabularer.

Det er med andre ord feltene som det er mulig å standardisere, men som bare til en viss grad er standardisert, som forårsaker de negative utslagene på oppgavens forskningsspørsmål. Manglende valg av autorisert navneform på *Creator* hemmer gjenfinning og interoperabilitet, mens manglende kontrollert vokabular for *Keyword* og *Type* påvirker både graden av gjenfinnbarhet, interoperabilitet og muligheten for gjenbruk. Det er flere årsaker til dette problemet. For det første er metadata-skjemaet lagt til rette med fritekst i stedet for faste verdier/vokabular for disse feltene. Dernest kan data-kurateringen være sviktende, ved at kurator ikke kvalitetssikrer forskerens utfylling av metadata-skjema.

Konklusjonen til Kim et al. (2019) er treffende også for min undersøkelse:

> The findings suggest that the metadata requested during data deposition should be defined in a more standardized and consistent manner, particularly those that can have various meanings within and across disciplines. This can help depositors better understand what metadata and documentation is required to prepare for deposition and assist curators to resolve inconsistencies. The use of standardized terminology can also

> support the interoperability of metadata across repositories, an innovation that would facilitate interdisciplinary data re-users' research. (s. 865)

Her er stikkordet *standardisering*; noe som vil hjelpe både metadata-skaper og forskeren som skal gjenbruke forskningsdataene. Et metadata-skjema med standardiserte verdier vil gjøre det enklere for forskeren å deponere sine data, og enklere for kuratoren å korrigere ved behov. Interoperabiliteten mellom ulike arkiver vil bli styrket, og dermed også muligheten for gjenbruk av forskere som representerer andre disipliner – både fordi forskningsdata blir tilgjengeliggjort, og fordi et standardisert vokabular gjør det lettere å søke opp og forstå dataene.

Balatsoukas et al. (2018) konkluderer også sin undersøkelse med å sette fokus på behovet for *standardisering*, og da særlig når det gjelder kontrollerte emneord. De påpeker at kontrollerte vokabularer vil bidra til økt overensstemmelse, samt en effektivisering av gjenfinnings- og gjenbruksprosessen (s. 7).

Målene for deling og gjenbruk av forskningsdata er klare, uttrykt gjennom FAIR-prinsippene. Utfordringen er hvordan man skal nå dem. Hodson et al. (2018) har følgende svar:

> This vision cannot be realised without specifications and standards for common components to enable interoperability across the FAIR data ecosystem. In addition to implementing the core concept of the FAIR Digital Object, two areas of activity have particularly high priority: 1) the development, refinement and adoption of shared vocabularies, ontologies, metadata specifications and standards which are central to interoperability and reuse at scale; 2) the increased provision and professionalisation of data stewardship, data repositories and data services. The first of these requires more concerted, coordinated and better resourced community efforts. (s. 12)

Her gis et todelt svar på hvordan oppnå FAIR-prinsippene. Den første delen berører metadata, og igjen er svaret *standardisering.* Den andre delen har med metadata-skaper og -kurator å gjøre, samt infrastrukturen for forskningsdata. Hodson et al. (2018) påpeker her viktigheten av å møte utfordringene med manglende standardisering på en samordnet og godt planlagt måte.

7.2 Vurdering av undersøkelsen

Formålet med denne oppgaven var å kartlegge kvaliteten på norske metadata for forskningsdata. Etter min vurdering har undersøkelsen avdekket klare trekk ved et representativ utvalg av metadata. Selv om datasettet ikke er veldig omfangsrikt, representerer det åtte ulike norske forskningsinstitusjoner. Det at resultatene viser en klar tendens, er også et argument for at utvalget er stort nok. Det er derfor legitimt å konkludere med at formålet med undersøkelsen – å vurdere kvaliteten på norske metadata for forskningsdata – er nådd.

7.3 Vurdering av analyseverktøyet

Denne oppgavens analyseverktøy ble utarbeidet med den hensikt å kunne analysere det enkelte element i en metadata-post opp mot hensiktsmessige kvalitetskrav. Det hadde vært mulig å kun bruke FAIR-prinsippene som vurderingskriterium for kvalitet, men for å få et bredere teoretisk fundament valgte jeg å se til både litteratur om metadata-kvalitet generelt, samt den hittil begrensede mengden tidligere forskning om kvaliteten på metadata for forskningsdata. Det viste seg å være en klar sammenheng mellom kvalitetskravene i henholdsvis FAIR-prinsippene og litteraturen, noe som ga analyseverktøyet en større legitimitet.

Ved å gå i detalj på et område, vil man alltid stå i fare for å miste overblikket og helheten. Dette er også en reell mulighet i denne analysen. Som sitert tidligere i oppgaven, hevder Hodson et al. (2018) at "notions of findability, accessibility, interoperability and reusability - and the actions needed to enable them - are so deeply intertwined that it does not make sense to address them individually" (s. 10). Dette blir imidlertid uttrykt i en kontekst som fokuserer på hvordan FAIR-prinsippene kan settes ut i livet gjennom endring av kultur og infrastruktur innenfor forskningsmiljøene. Som en del av et analyseverktøy kan FAIR, etter min oppfatning, deles opp i sine enkelte faktorer og brukes som et mål på metadata-kvalitet.

Å analysere de metadatafeltene som er blitt undersøkt i tidligere forskning, og som har innvirkning på kravene som fremkommer både i FAIR-prinsippene og i litteraturen om metadata-kvalitet for øvrig, har etter min mening fungert tilfredsstillende. At resultatene fra denne undersøkelsen samsvarer med andre undersøkelser av kvaliteten på metadata for forskningsdata, er dessuten en klar indikasjon på at analyseverktøyet har vært velegnet.

7.4 Perspektivering

Denne undersøkelsen indikerer at norske metadata for forskningsdata har høy kvalitet på noen områder og lav kvalitet på andre områder, og at resultatene i stor grad samsvarer med tidligere forskning. Da dette er den første undersøkelsen i sitt slag på norske forhold, kan den bidra til økt fokus på viktigheten av høy kvalitet på metadata for forskningsdata ved våre forskningsinstitusjoner. Willis et al. (2012) argumenterer for at "more metadata-focused research is needed to move toward an increasingly interoperable environment where scientific data is to be shared across domains and communities, and existing data silos are eliminated" (s. 1508).

Undersøkelsen viser at også i Norge er mer forskning nødvendig på området kvaliteten på metadata for forskningsdata, og at fokus må rettes mot tiltak som øker interoperabiliteten og muligheten for gjenbruk – nemlig standardisering i form av vokabularer og forbedrede metadata-skjemaer. Dette for å eliminere henholdsvis semantisk og strukturell heterogenitet i størst mulig grad (Haslhofer & Klas, 2010, s. 14-16). Som nevnt innledningsvis, er det ekstra viktig å bruke et standardisert vokabular ved lagring av forskningsdata, selv om det er viktig i alle digitale arkiver. Dette fordi datasett kan brukes flere ganger og resultere i en rekke ulike versjoner. Det er da avgjørende å gjøre bruk av kontrollerte vokabularer, slik at de semantiske relasjonene mellom datasett og innenfor livssyklusen til et spesifikt datasett blir ivaretatt (Balatsoukas et al., 2018, s. 7).

I følge Zhang et al. (2015) er forskning med fokus på funksjonalitet som støtter kontrollerte emneord for forskningsdata nærmest fraværende. For å styrke dette området må både vokabularer utvikles i samarbeid med forskerfellesskapene, og metadata-skjema som legger til rette for faste vokabularer må utvikles. For generiske arkiver er det optimale muligheten for tilgang til flere ulike vokabularer fra ett og samme skjema. Både Zhang et al. (2015) og Karimova (2018) presenterer undersøkelser som sikter mot å utvikle nettopp metadata-modeller for forskningsdata som legger til rette for "multiple controlled vocabularies" (Zhang et al., 2015, s. [7]). En slik løsning vil gjøre det lettere for forskeren å deponere sine data, samtidig som kvaliteten på metadata vil styrkes. Selv om "cost, usability, and other challenges surface when trying to work with more than one vocabulary in a single system" (Zhang et al., 2015, s. [1]), er det antagelig den beste veien å gå for å løse utfordringen med mangelfull standardisering av emneord og datatype-betegnelser.

Kunnskapsdepartementet (2017) legger vekt på viktigheten av at verdien i forskningsdata utvunnet ved offentlig finansiering må utnyttes best mulig. Da er det ikke tilstrekkelig med åpen og sikker lagring av dataene. De må også være reelt gjenfinnbare og gjenbrukbare, og FAIR-prinsippene viser hvordan dette kan oppnås blant annet ved å sørge for høy kvalitet på metadataene (s. 25). Det er med andre ord også i Regjeringens interesse at kvaliteten på metadata for forskningsdata blir forsket videre på og forbedret. Avslutningsvis er det derfor min anbefaling at fokus rettes mot mer forskning og utvikling på området norske metadata for forskningsdata.

Litteratur

Austin, C. C., Brown, S., Fong, N., Humphrey, C., Leahey, A. & Webster, P. (2016). Research Data Repositories: Review of Current Features, Gap Analysis, and Recommendations for Minimum Requirements. *IASSIST Quarterly*, *39*(4), 24. https://doi.org/10.29173/iq904

Balatsoukas, P., Rousidis, D. & Garoufallou, E. (2018). A method for examining metadata quality in open research datasets using the OAI-PMH and SQL queries: the case of the Dublin Core 'Subject' element and suggestions for user-centred metadata annotation design. *International Journal of Metadata, Semantics and Ontologies*, *13*(1), 1-8. https://doi.org/10.1504/IJMSO.2018.096444

Barton, J., Currier, S. & Hey, J. (2003). *Building Quality Assurance into Metadata Creation: An Analysis based on the Learning Objects and e-Prints Communities of Practice*. Innlegg presentert ved International Conference on Dublin Core and Metadata Applications. Abstract hentet fra https://dcpapers.dublincore.org/pubs/article/view/732/728

Bruce, T. R. & Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting. I D. I. Hillmann (Red.), *Metadata in practice* (1. utg., s. 238-256). Chicago, IL: American Library Association.

Caplan, P. (2003). *Metadata fundamentals for all librarians*. Chicago: American Library Association.

Choudhury, S., Cowles, E., Croft, H., Estlund, K., Fary, M., Faustino, G., … Waters, N. (2018). *Research Data Curation A Framework for an Institution-Wide Services Approach*.

DataCite Metadata Working Group. (2019). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3. DataCite e.V. Hentet fra https://schema.datacite.org/meta/kernel-4.3/doc/DataCite-MetadataKernel_v4.3.pdf

DataverseNO. (u.å.). Arkiver dataa dine. Hentet fra https://site.uit.no/dataverseno/nn/arkivering/arkiver-dataa-dine/

Day, M. (2010). *DCC Digital Curation Manual: Instalment on Metadata* HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.

Farnel, S. & Shiri, A. (2014). Metadata for Research Data: Current Practices and Trends. *Dublin Core Conference* (s. 74-82).

G8 Science Ministers. (2013). G8 Science Ministers Statement. Hentet fra https://www.gov.uk/government/news/g8-science-ministers-statement

Gilliand, A. J. (2008). Setting the stage. I M. Baca (Red.), *Introduction to metadata* (2. utg., s. 1-19). Los Angeles, CA: Getty.

GO FAIR. (u.å.). FAIR Principles. Hentet fra https://www.go-fair.org/fair-principles/

Gordon, K. (2013). *Principles of data management : facilitating information sharing* (2. utg.). Swindon: BCS Learning & Development Limited.

Greenberg, J., White, H. C., Carrier, S. & Scherle, R. (2009). A Metadata Best Practice for a Scientific Data Repository. *Journal of Library Metadata, 9*(3-4), 194-212. https://doi.org/10.1080/19386380903405090

Haslhofer, B. & Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys (CSUR), 42*(2), 1-37. https://doi.org/10.1145/1667062.1667064

Hider, P. (2018). *Information Resource Description: Creating and managing metadata*. United Kingdom: Facet Publishing.

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., … Wittenburg, P. (2018). Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data. https://doi.org/10.5281/ZENODO.1285271

Karimova, Y. (2018). Flexible metadata models and controlled vocabularies for research data description in multiple domains. *Dublin Core 2018 Doctoral Consortium*.

Kim, J., Yakel, E. & Faniel, I. M. (2019). Exposing Standardization and Consistency Issues in Repository Metadata Requirements for Data Deposition. *College & Research Libraries, 80*(6), 843-875.

Kunnskapsdepartementet. (2017). *Nasjonal strategi for tilgjengeliggjøring og deling av forskningsdata*. Hentet fra https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-tilgjengeliggjoring-og-deling-av-forskningsdata/id2582412/

Lewis, G. A., Morris, E., Simanta, S. & Wrage, L. (2008). Why Standards Are Not Enough to Guarantee End-to-End Interoperability. *Seventh International Conference on Composition-Based Software Systems (ICCBSS 2008)* (s. 164-173). https://doi.org/10.1109/ICCBSS.2008.25

Marc, D. T., Beattie, J., Herasevich, V., Gatewood, L. & Zhang, R. (2016). Assessing Metadata Quality of a Federally Sponsored Health Data Repository. *AMIA ... Annual Symposium proceedings. AMIA Symposium,Medline, 2016*, 864-873.

Miller, P. (2000). Interoperability. What Is It and Why Should I Want It? Hentet fra http://www.ariadne.ac.uk/issue24/interoperability/

Mons, B., Neylon, C., Velterop, J., Dumontier, M., Da Silva Santos, L. O. B., Wilkinson, M. D. & Bioinformatics. (2017). Cloudy, increasingly FAIR; Revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services and Use*, *37*(1), 49-56. https://doi.org/10.3233/ISU-170824

NISO. (2004). *Understanding Metadata*. Bethesda, MD: NISO.

NSD. (u.å.). Hva er en datahåndteringsplan? Hentet fra https://nsd.no/arkivering/hva_er_en_datahandteringsplan.html

Ochoa, X. (2014). Metadata quality. I *Handbook of Metadata, Semantics and Ontologies* (s. 63-88). World Scientific Publishing Co. Pte. Ltd.

OECD. (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris: OECD Publishing.

RDA-Codata Legal Interoperability Interest Group. (2016). Legal Interoperability Of Research Data: Principles And Implementation Guidelines. I.

Riley, J. (2017). *Understanding Metadata: What is Metadata, and What is it For?* Baltimore: National Information Standards Organization.

Rousidis, D., Garoufallou, E., Balatsoukas, P. & Sicilia, M. A. (2014). Metadata for big data: a preliminary investigation of metadata quality issues in research data repositories. *Information Services & Use*, *34*(3-4), 279-286. https://doi.org/10.3233/ISU-140746

Rousidis, D., Garoufallou, E., Balatsoukas, P. & Sicilia, M. A. (2015). Evaluation of metadata in research data repositories: The case of the DC.Subject element. I Garoufallou E., Hartley R. & G. P. (Red.), *Metadata and Semantics Research. MTSR 2015. Communications in Computer and Information Science* (bd. 544, s. 203-213). Springer, Cham.

Tani, A., Candela, L. & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. *Information Processing and Management*, *49*(6), 1194-1205. https://doi.org/10.1016/j.ipm.2013.05.003

UIO. (2015). *Dataeksplosjonen – en stor utfordring, og en gedigen mulighet!: rapport fra arbeidsgruppen «Lagring og deling av forskningsdata»* Oslo. Hentet fra

https://www.usit.uio.no/om/organisasjon/itf/saker/forskningsdata/lagring-og-deling-av-forskningsdata.pdf

UNC University Libraries. (2019). Metadata for Data Management: A Tutorial. Hentet fra https://guides.lib.unc.edu/metadata/standards

University of Vienna. (u.å.). Research Data Management. Hentet fra https://datamanagement.univie.ac.at/en/research-data-management/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3:160018*. https://doi.org/10.1038/sdata.2016.18

Willis, C., Greenberg, J. & White, H. (2012). Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology*, *63*(8), 1505-1520. https://doi.org/10.1002/asi.22683

Yasser, C. M. (2011). An Analysis of Problems in Metadata Records. *Journal of Library Metadata*, *11*(2), 51-62. https://doi.org/10.1080/19386389.2011.570654

Zhang, Y., Ogletree, A., Greenberg, J. & Rowell, C. (2015). Controlled vocabularies for scientific data: Users and desired functionalities. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-8. https://doi.org/10.1002/pra2.2015.145052010054