Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

Brage Vejlgaard Sørensen

Datadeling i forskningsmiljøer

En empirisk casestudie

Masteroppgave i Digital samhandling
Veileder: Elena Parmiggiani
Medveileder: Nana Kwame Henebeng Amagyei
Juni 2023

NTNU
Kunnskap for en bedre verden

Brage Vejlgaard Sørensen

Datadeling i forskningsmiljøer

En empirisk casestudie

Masteroppgave i Digital samhandling
Veileder: Elena Parmiggiani
Medveileder: Nana Kwame Henebeng Amagyei
Juni 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for datateknologi og informatikk

NTNU
Kunnskap for en bedre verden

Sammendrag

Den teknologiske fremgangen har bidratt med å gi store muligheter og skape nytekning når det kommer til hvordan forskere henter inn data, behandler denne og hvilke roller personer kan ha i forskning. Rollen til den tradisjonelle forskeren er endret og de må nå bruke nye verktøy og metoder for å hente inn, behandle, beskrive data og dele. De nye mulighetene fører samtidig med seg utfordringer knyttet til store mengdene data som genereres. Hvordan kan forskerne stole på kvaliteten på all dataene som genereres og hvordan kan de dele alle dataene med hverandre?

Forskningen i dette prosjektet har som mål å bidra med et empirisk bidrag som beskriver hvordan forskere bruker teknologier og verktøy til å hente inn, behandle, beskrive og dele data med hverandre. Samtidig skal forskningen være med å gi et teoretisk bidrag til forskere og andre som behandler data ved å beskrive problemet knyttet til hvordan databehandling i tidligere forskning blir beskrevet som en perfekt strømlinjeformet prosess uten store utfordringer. Prosjektet skal også være med å besvare hvordan forskere kan dele data på en mer effektiv måte og hvordan forskere kan beskrive data for å gjøre de mer forståelig.

For å finne svar på spørsmålene og kunne gi et bidrag til andre er det gjennomført en empirisk casestudie. Studien er har en interpretivistisk tilnærming og datainnsamlingen er gjennomført med observasjoner og ustrukturerte intervjuer. Funnene i oppgaven er basert på følelsene og holdningene som kommer frem i intervjuene og hvordan forskerne jobber med data basert fra observasjonene. Informantene som har vært en del av studien er forskere eller dataforskere som arbeider med data.

Funnene viser hvordan det er store utfordringer og muligheter når det kommer til databehandling i de undersøkende forskningsmiljøene. Databaser og datadepoter benyttes til å dele data med andre selv om det er knyttet utfordringer til hvordan gjennomføringen skjer i praksis. Det er også utfordringer knyttet til hvordan forskerne beskriver dataen de besitter. Til slutt kommer det frem hvordan databehandling ikke er like strømlinjeformet som beskrevet av tidligere forskning, men heller er en mer kompleks og kontinuerlig prosess som må tas til etterretning.

Abstract

Technological progress has contributed to providing great opportunities and creating innovation when it comes to how researchers collect data, process data and what roles people may have in research. The role of the traditional researcher has changed, and they must use new tools and methods to collect, process, describe, and share their data. The new possibilities also bring challenges related to enormous amounts of data that are generated. How can researchers trust the quality of all the data generated and how can they share all the data with each other?

The research in this project aims to make an empirical contribution that describes how researchers use technologies and tools to collect, process, describe and share data with each other. At the same time, the paper will help make a theoretical contribution to researchers and others who process data. The paper will help address this problem by contradicting how former research present data as a perfectly streamlined process without major challenges. The project will also help to answer how researchers can share data in a more efficient way and how researchers can describe data to make it more understandable.

To find answers to the questions and be able to contribute to others, a multiple case study has been conducted. The study has an interpretive approach, and the data collection is conducted with observations and unstructured interviews. The findings in the thesis are based on the feelings and attitudes that emerge in the interviews, and how the researchers work with data based on the observations. The informants who have been part of the study are researchers or computer scientists who work with data.

The findings show how there are major challenges and opportunities when it comes to data processing in investigative research environments. Databases and data repositories are used to share data with others, although there are challenges associated with how the implementation takes place in practice. There are also challenges related to how researchers describe the data they possess. Finally, it emerges how data processing is not as streamlined as described by previous research. However, it is a more complex and continuous process that must be considered.

Forord

Denne masteroppgaven er gjennomført ved Institutt for datateknologi og informatikk ved Norges teknisk-naturvitenskapelig universitet (NTNU) og marker avslutningen for det toårige masterstudiet i Digital Samhandling i Trondheim.

Oppgaven er gjennomført som en empirisk casestudie og jeg ønsker å takke begge caseorganisasjonene som har stilt opp og gjort dette mulig. En spesiell takk rettes til de to nøkkelpersonene fra casen om industriell forskning. Dere har vært med å forme oppgaven og retningen den tok hadde ikke vært mulig uten dere. Samtidig ønsker jeg å takke nøkkelpersonen i den økologiske casen som gjorde det mulig å besøke deres arbeidsplass og la til rette for intervjuer. Dere tre vet hvem dere er, og det hadde vært umulig å gjennomføre uten dere.

Jeg vil også takke alle informanter som har deltatt i undersøkelsen og gitt av sin arbeidstid for å gjøre dette mulig.

Videre ønsker jeg å takke alle medstudenter for to morsomme år på studiet og alle andre som har gjort dagene av og på campus overkommelige. Jeg hadde ikke vært her jeg er i dag uten noen av dere.

Til slutt ønsker jeg å takke veileder Elena Parmiggiani og biveileder Nana Kwame Henebeng Amagyei. Dere har hjulpet meg gjennom oppgaven og jeg har satt stor pris på alle gode råd, veiledning og diskusjoner.

Trondheim

Mai, 2023

_____

Brage Vejlgaard Sørensen

Innhold

Figurer

Tabeller

Forkortelser/symboler

SDI – Stegvis-deduktiv induksjon
PLOS – Public Library of Science
EML – Ecological Metadata Language

1 Introduksjon

Det genereres store mengder data hver dag. Dataen som genereres kommer fra enkeltpersoner som bruker digitale verktøy for å løse arbeidsoppgaver på jobb eller skole, leser nettavisen på mobilen eller gjennomfører oppgaver som påvirker et aspekt en forsker ønsker å utforske. All denne dataen kan måles på en eller annen måte, noe som skaper måter å beskrive tidligere fenomener eller predikere hendelser som kan oppstå. Samtidig eksisterer det mange verktøy som beskriver hva dataen faktisk betyr, hva den kan brukes til og hvordan den har oppstått. Denne beskrivelsen, eller metadataen, er utrolig viktig for å skape en forståelse for konteksten dataen tilhører(Alves et al., 2018; Borer et al., 2009).

Et annet aspekt med dataene som kommer inn er hvordan forskere som ønsker å bruke denne dataen kan forsikre seg om kvaliteten på dataen. Hvordan kan forskere vite dataen de har fått tak i enten gjennom egne undersøkelser, fra andre forskere eller brukergenerert data består av høy kvalitet og kan være med å skape gode resultater i forskningen? Disse spørsmålene er små innen databehandling, men er utrolig viktige for data sin kontekst og validitet (Cai & Zhu, 2015). I denne oppgaven er det ønskelig å bringe denne utfordringen frem til forskere og forskningsinstitusjoner. Ved å trekke dette aspektet fra den strømlinjeformede databehandlingsprosessen beskrevet i tidligere forskning (Bordelon, 2023), vil det være mulig å belyse hvor viktig dette temaet er for forskning. Samtidig vil det være mulig å argumentere imot den teoretiske fremstilling av databehandling som en strømlinjeformet prosess hvor de forskjellige fasene kommer etter hverandre og heller se på databehandling som et helhetlig hjul hvor hver fase har påvirkning på hverandre og går i sirkel.

Videre vil oppgaven se på hvordan forskere deler data. Med den teknologiske fremgangen er det ikke bare metoder for datainnsamling som har endret seg. Muligheten forskere har til å samarbeide og dele data har gjennomgått store endringer de siste årene. Databaser og datadepoter som benytter skylagring og tilgangsstyring gir store muligheter når det kommer til deling av rådata og prosesserte data (Allagnat et al., 2019). På en annen side er det utfordringer knyttet til den sosiale utfordringen ved å dele data (Karasti et al., 2006). Derfor vil dette aspektet også utforskes og besvares i oppgaven.

Målet med denne oppgaven er å studere hvordan forskere arbeider med databehandling i ulike forskningskontekster. Gjennom observasjoner med ustrukturerte intervjuer skal datainnsamlingen danne et grunnlag til å sammenligne praksisen til forskere i de forskjellige miljøene for å identifisere likheter og forskjeller. Dermed vil det være mulig å skape og diskutere en forståelse av hvordan vitenskapelige data håndteres i ulike miljøer med ulike forutsetninger.

Videre skal prosjektet være med å løse en utfordring for forskere. Mye dataarbeid blir presentert som en strømlinjeformet prosess der alt er perfekt. Det viser seg databehandlingsprosessen ikke er like perfekt som tidligere forskning beskriver den. Dermed skal oppgaven være med å belyse denne utfordringen for forskere som behandler data og for forskere som forsker på andre som behandler data.

Oppgaven vil være med å gi et empirisk bidrag og et teoretisk bidrag til forskning innenfor databehandling i forskningsmiljøer. Det empiriske bidraget vil være en beskrivelse av hvordan forskere jobber med forskningsdata i forskjellige miljøer, beskrivelser av hvorfor de velger den fremgangsmetoden de har og data som beskriver hvilke systemer og verktøy forskerne bruker. Det teoretiske bidraget er med å belyse en utfordring for alle forskere som jobber med databehandling. Det teoretiske problemet som oppgaven prøves å besvare er hvordan forskere ser på databehandling som en prosess og hvilke endringer i atferd forskere burde ha knyttet til hvordan de beskriver databehandling. Samtidig kan det teoretiske bidraget være med å påvirke hvordan databehandlere som ikke er forskere tenker om databehandling

Dette forskingsprosjektet er et planlagt prosjekt som tar utgangspunkt i forskere og hvordan de bruker digitale verktøy til å samle, prosessere, validere og dele data. Før oppstart er det gjennomført et pilotprosjekt der målet var å danne et grunnlag og en innsikt i hvilke utfordringer som er eksisterende i forskningsmiljøene og en bredere forståelse gjennom tidligere forskning og litteraturgjennomgang.

Med utgangspunktet i formålet til prosjektet og bakgrunn for gjennomføringen er det utarbeidet to forskningsspørsmål.

F1: Hvordan kan forskere gjøre tilgang til forskningsdata enklere?
F2: Hvordan kan forskere gjøre forskningsdata mer forståelig?

Forskningsspørsmålene vil være med å besvare formålet til prosjektet og konkretisere noen av de aktuelle utfordringene forskerne står ovenfor. Det første forskningsspørsmålet handler om hvordan forskere deler data med hverandre i dag og hvilke tiltak eller endringer av atferd de kan ha for å gjøre denne prosessen enklere. Det andre forskningsspørsmålet beskriver en annen konkret utfordring forskere står ovenfor. Hvordan skal de beskrive dataen de deler slik at andre forstår de? Det er ikke nok å bare dele data med hverandre, forskerne må også forstå hva dataene de får tak i betyr og hva de kan brukes til.

Videre i oppgaven beskrives først den teoretisk bakgrunnen og den tidligere forskningen som er gjort på området. Deretter blir de to casene presentert. I denne beskrivelsen blir viktige aspekter ved de forskjellige casene presentert og hvordan de henger sammen. Videre beskrives metodekapittelet. I dette kapittelet beskrives hvordan studien er gjennomført, begrunnelse for forskjellige valg som er gjort og kapittelet avsluttes med en egenvurdering av metodevalg og gjennomføring. Etter metoden blir de analyserte resultatene presentert etterfulgt av en diskusjon. Diskusjonen er todelt hvor det casene blir sammenlignet i den første delen, etterfulgt av en del hvor funnene fra sammenligningen blir diskutert opp mot tidligere forskning. Oppgaven avsluttes med en konklusjon hvor forskningsspørsmålene bli besvart og det blir nevnt begrensninger med oppgaven, hva som burde gjøres videre og hvordan prosjektet relaterer til bærekraft.

2 Teoretisk bakgrunn

Dette kapittelet inneholder litteraturbakgrunnen og beskriver et overordnet teoretisk rammeverk for oppgaven. Databehandling innen forskning blir ansett på forskjellige måter. En tilnærming for databehandling er en strømlinjet inndeling i fire hvor databehandling er prosessen hvor data innhentes og lagres, prosesseres, valideres, og deles (Bordelon, 2023). En annen tilnærming er å dele i fem hvor databehandling blir ansett som innsamling, datakurering, forbruk, konseptualisere og kontroll (misqresearchcurations, 2022). Ved slå sammen konseptene fra de to ståstedene er det i denne studien valgt å se på databehandling som innsamling, lagring og prosessering, kvalitetssikring, og deling. Gjennomgangen av det teoretiske grunnlaget starter med å beskrive delprosessen for datainnsamling. Deretter beskriver den delprosessen for datalagring og prosessering før delprosessen for kvalitetssikring blir gjennomgått. Den siste delprosessen for databehandling som blir gjennomgått er deling av data før den teoretiske bakgrunnen avsluttes ved å gjennomgå noen byråkratiske regler med tanke på eierskap til data og en oppsummering.

2.1 Delprosess for datainnsamling

Datainnsamling er en databehandlingsaktivitet med fokus på å identifisere egnede datakilder, designe datainnsamlings protokoller og etablere kvalitetskontroller over datainnsamlingen og innhenting av data (misqresearchcurations, 2022). Tidlig ble det etablert utfordringer ved datainnsamling som gikk på anvendelse av teknologi og det har senere vært en utvikling som muliggjør datainnsamling på en helt annen måte enn tidligere (misqresearchcurations, 2022).

Fremveksten av internett, teknologiske løsninger for datainnsamling og metadataløsninger for å beskrive datainnsamlingen ga forskere muligheter til å hente data med nye metoder (Alves et al., 2018; misqresearchcurations, 2022). Internett ga forskerne muligheten til å generere enorme datasett. De enorme data mengdene er det som er ansett som Big Data (De Mauro et al., 2015). Et økende volum av brukergenerert data gir forskerne et større datagrunnlag til sine prosjekter. På en annen side vil det være utfordringer knyttet til kvaliteten på dataen, og forskere må i større grad vurdere i hvilken grad data er av kvalitet og hvilke utfordringer som eksisterer knyttet til å sikre høy kvalitet (Diepenbroek et al., 2014; misqresearchcurations, 2022).

Utviklingen av teknologier og metoder for å samle inn data, drev også en parallell utvikling om rollene knyttet til datainnsamling (misqresearchcurations, 2022). Hvor forskere tidligere gjorde mye av arbeidet selv, gir den teknologiske utviklingen muligheter for programmerere og ingeniører å være delaktige i datainnsamlingsprosessen ved å trekke ut data fra eksterne kilder. Vanlige mennesker ble informanter ved å bruke teknologiske løsninger og forskere kunne gjenbruke eksisterende data i større grad (misqresearchcurations, 2022).

Denne utviklingen har også ført til fremskritt og muligheter for hvordan forskere henter inn data i dag. Dataen kan være alt fra fjernmålinger og nyeste generasjons sekvenseringsdata til feltobservasjoner eller innsamlinger (Diepenbroek et al., 2014). Denne dataen kan innhentes automatisk gjennom dataloggere eller samles individuelt av

forskere. Resultatet av disse datainnsamlingene vil gi data av varierende kvalitet i forskjellige datapakker. Samtidig kan metadataene som samles inn være begrenset, noe som er en utfordring (Diepenbroek et al., 2014).

2.2 Delprosess for datalagring og prosessering

Datalagring og prosessering er en databehandlingsaktivitet som innebærer infrastrukturer som gjør det mulig å samle inn, beskrive, lagre og analysere data. Aktiviteter som må gjennomføres er å lage et datalagringssystem, rense og transformere data til et format som kan lagres i datalagringssystemet, kategorisere og beskrive data samt gjennomføre sikkerhetstiltak med tanke på lagring (Borer et al., 2009; Diepenbroek et al., 2014; misqresearchcurations, 2022).

Vanligvis produseres forskningsdata innen økologi i regneark eller lokale databaser som er spesialiserte opp mot det spesifikke forskningsområdet (Diepenbroek et al., 2014). Denne lagringsmåten er ikke optimal eller hensiktsmessig for forskningen. Det er vanskelig å administrere komplekse datasett, spesielt hvis datasettet har tilknytting til andre, eksisterende, datasett eller er multimediadata. Dette gjenspeiler seg hos Borer, som mener: «Excel and Access have not always been the prefered spreadsheet/data storage programs, and in the future they will, undoubtedly, be replaced by others (Borer et al., 2009).» En annen utfordring knyttet til datainnsamling i økologi er innsamling av data i felt. Verktøy for lagring av data må være fleksibelt og brukbar i felt uten internettilgang, men også være mulig å integrere i forskningsmiljøet gjennom forskjellige nettverksressurser (Diepenbroek et al., 2014).

Neste aktivitet som inngår i datalagring og prosessering er beskrivelse av data eller metadata. «Metadata er informasjon som beskriver annen informasjon, altså data om data (Gjersdal & Nätt, 2023).» Metadataen vil dermed brukes til å enklere finne igjen dataen eller gjøre den mer lesbar ved eventuell deling. For å beskrive dataen blir det foreslått å ha metadata som beskriver tid og sted for innsamling, tid for opprettelse i lagringssystemet og annen informasjon som gir et godt innblikk i hva dataen egentlig betyr (Borer et al., 2009; Gjersdal & Nätt, 2023).

Corvallis_VegBiodiv_2007.csv

Date	Site	SpName	Abundance
2007	XYZ	Sp. 1	21
2007	XYZ	Sp. 4	45
etc…			

Figur 1: (Borer et al., 2009) Forslag til hvordan langsiktig datalagring kan se ut i regneark.

Borer et al., (2009) mener Figur 1 er et godt eksempel på hvordan langsiktig datalagring kan se ut. Metadataen beskriver kritisk informasjon om dataen i hver kolonne istedenfor å være en del av filnavnet. Samtidig påpekes viktigheten av å alltid opprettholde

effektive metadata eller vurdere å bruke Ecological Metadata Language (EML) for å guide hvilke metadata som er viktig for forskningen (Borer et al., 2009).

Videre viser forskningen hvordan 40 forskjellige metadataverktøy er brukt eller nevnt i vitenskapelige manuskript fra 1997, som er den første gang dette er nevnt, til 2018 (Alves et al., 2018). Selv om majoriteten av disse verktøyene er skjemaer og standarder, er det også metadataverktøy som metadatakataloger, ontologier og utvidelser til metadataredaktører og kodingsstandarder. EML er en av metastandardene som er mye brukt siden implementering i 1997 og dermed blir Borer et al. sitt forslag om å bruke den standarden styrket av denne forskningen (Alves et al., 2018; Borer et al., 2009).

Utfordringen med alle initiativene til økt metadataverktøy og plattformer er hvordan forskerne skal bruke dem, spesielt innen økologi. Alves et al. mener et økende initiativ med tekniske løsninger vil føre til en mangel på metadatamodeller som beskriver de ulike datatypene (Alves et al., 2018). Tidligere har forskere beskrevet datasett innenfor konteksten til prosjektavtalen og etter en valgt datamodell. Hvilken datamodell som passer til hvilket forskningsprosjekt og å følge den valgte datamodellen er en utfordring i seg selv. Derfor vil dette bli enda mer komplekst om de nye løsningene ikke kan integreres metadatastandaren eller metadatamodellen forskerne ønsker å bruke.

En annen aktivitet er å renske dataen og transformere dataen til et filformat som er tilpasset lagringssystemet. Som nevnt tidligere kan Excel og Access være på vei ut og nye og potensielt uforenlige systemer med dataformatene kan være på vei inn (Borer et al., 2009). De mener også det samme gjelder for maskinvare og at formater kan bli raskt foreldet. Faren med foreldede systemer er tap av verdifulle data fordi det ikke er kompatibelt med de teknologiske vinningene. Derfor anbefales det å lagre data i det beste «ikke-proprietære formatet» som vil være internett (Borer et al., 2009).

Den siste aktiviteten er analyse av data. Borer et al., anbefaler å bruke et skriptbasert analyseprogram for å analysere dataen. Analyseskript er skriftlige registreringer av de ulike trinnene som er involvert i databehandling og dataanalyse og kan være med å gi en analytisk form for «metadata» (Borer et al., 2009). Skriptene kan enkelt revideres og kjøres i flere omganger avhengig av hva forskeren er ute etter eller hvilke tilbakemeldinger som blir gitt. Andre fordeler med å bruke skript for analyse er muligheten til å ha oversikt over hva som ble gjort med dataene fra innsamlingstidspunktet til publisering eller deling (Borer et al., 2009). Skript kan også brukes på flere måter og resultatene kan enkelt visualiseres i forskjellige former som grafer, tabeller eller andre (Vliet, 2019).

2.3 Delprosess for kvalitetssikring

Det eksisterer flere måter å beskrive hva datakvalitet er for noe. Professor Richard Y. Wang ved MIT har gjennomført forskning på området og beskriver datakvalitet som bruksegnet (Wang & Strong, 1996). Samtidig definerte de datakvalitetsdimensjoner til å være et sett av datakvalitetsattributter som representerer et aspekt eller en konstruksjon av datakvalitet. Videre har de identifisert fire kategorier som inneholder femten datakvalitetsdimensjoner (Wang & Strong, 1996).

En annen beskrivelse om prinsippene rundt datakvalitet er gjennomført i USA. De har delt disse prinsippene inn i tre forskjellige prinsipper.

1. Data er et produkt, med kunder, som de har kostnad og verdi for.
2. Som et produkt, har data kvalitet, et resultat av prosessen der data genereres.

3. Datakvalitet avhenger av flere faktorer, inkludert formålet dataen brukes til, bruker og tid for generering (Cai & Zhu, 2015).

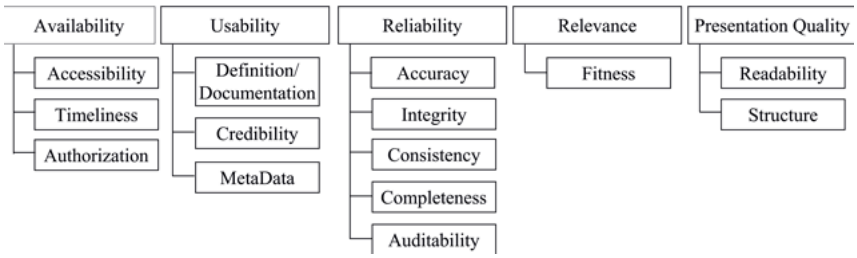
Samtidig administreres ofte forskningsdata av enkeltforskere uten støtte fra andre IT-eksperter (Diepenbroek et al., 2014). Derfor er dataene ofte dårlig strukturert og løsningene som brukes til å styre dataene er ikke hensiktsmessige. Forskerne bruker liten tid på datahåndtering, ofte fordi de er under press for å produsere publikasjoner med resultater. Dette fører til en lav prioritet til langsiktig håndtering av data. Som et resultat av dette kan det oppstå feil og hull i primærforskningsdata. Dette kan hindre gjenbruk av data og dataen blir dermed lite brukeregnet (Diepenbroek et al., 2014)

For å løse denne utfordringen blir data i biologisk- og økologiskforskning ofte arbeidet med av en dedikert dataansvarlig i større prosjekter. Den med rollen som dataansvarlig har ansvaret for å sikre data av høy kvalitet ved å manuelt sjekke data produsert av forskerne. Det eksisterer verktøy som støtter datakvalitetsstyring, men de er som oftest dårlig integrert i datainnsamlingsverktøyene. Med rollen som dataansvarlig får man dermed som oppgave å sikre brukeregnethet i dataen (Cai & Zhu, 2015; Diepenbroek et al., 2014).

En annen utfordring knyttet til kvalitet i data er knyttet til Big Data.

1. Mangfoldet av datakilder gir rikelig med datatyper og komplekse datastrukturer og øker vanskeligheten med dataintegrasjon.
2. Datavolumet er enormt, og det er vanskelig å bedømme datakvaliteten innen rimelig tid.
3. Data endres veldig raskt og aktualiteten til data kan være veldig kort, noe som krever høyere krav til prosesseringsteknologi.
4. Ingen enhetlige og godkjente datakvalitetsstandarder har blitt dannet i Kina eller i utlandet, og forskning på datakvaliteten til Big Data har akkurat begynt (Cai & Zhu, 2015).

I 2015 var Big Data et nytt konsept uten en uniform definisjon på datakvalitet eller datakvalitetskriterier. Ved å se på Big Data fra et forretningsperspektiv og hvilke datakvalitetsdimensjoner som var allmenn akseptert, ble det utarbeidet en to-lags modell som beskriver de forskjellige dimensjonene med elementer.



Figur 2 «A universal, two-layer Big data quality standard for assesment (Cai & Zhu, 2015)»

Figur 2 beskriver de fem dimensjonene med underliggende elementer. Tilgjengelighet omhandler om et datatilgangsgrensesnitt er gitt, om dataen enkelt kan gjøres offentlig eller kjøpes, om dataene kommer frem i tide innen en gitt tid, om dataene oppdateres regelmessig og om hvorvidt tidsintervallet fra datainnsamling og behandling til utgivelse møter de satte kravene (Cai & Zhu, 2015).

Videre handler brukervennlighet om dataen kommer fra spesialiserte organisasjoner i et land felt eller en bransje, om eksperter eller spesialister reviderer og kontrollerer dataene

regelmessig og om dataen er innenfor det akseptable omfanget for verdier (Cai & Zhu, 2015).

Den neste dimensjonen omhandler pålitelighet i dataene. I denne dimensjonen må dataene være nøyaktige, konsekvente, ha integritet og være komplette (Cai & Zhu, 2015).

Dataene som blir innhentet må også være relevante. De skal ikke nødvendigvis samsvar helt med teamet, men de må forklare et aspekt. Samtidig må dataene gi være i samsvar med brukerens søkningstema eller hentingstema (Cai & Zhu, 2015).

Til slutt må dataene være lesbare. Det skal være et format som er klart og forståelig, de skal være enkle å forstå og det skal være med en beskrivelse og klassifikasjon som tilfredsstiller brukerens behov (Cai & Zhu, 2015).

2.4 Delprosess for datadeling

En av metodene Public Library of Science (PLOS) benytter for å støtte åpen vitenskap og datadeling er ved bruk av en streng datatilgjengelighetspolicy som de etablerte tilbake i 2014 (Hrynaszkiewicz et al., 2021). Selv om PLOS, og andre, har introdusert forskjellige forslag til beste praksis for datadeling blir dette tatt i bruk av et mindretall av forskere. Det fortsetter å eksistere utfordringer knyttet til deling av forskningsdata og annen forskning begrunner dette i mangel på tid, ressurser, insentiver og/eller ferdigheter til å dele data (Hrynaszkiewicz et al., 2021).

I Japan, begrunnes deling av data med at dataene de finner kan være viktige for andre og motivasjonen for å dele data i grunner i mulighetene til å ha en økende fremgang i feltet og for å ha åpenhet og muligheter for gjenbruk av data (Allagnat et al., 2019). Selv om data deling er viktig i landet er det tydelig data ikke blir delt på en optimal måte. Av de som deler data, velger 35% å dele dataene privat og som oftest innenfor egen institusjon. Av respondentene i forskningen har 62% delt data privat og offentlig, 36% bare privat og kun 2% har bare delt dataene offentlig. For å dele dataene, brukes hovedsakelig tre metoder for privat deling og tre metoder for offentlig deling. Bruk av e-post, USB eller minnepenn og skyløsninger er mest bruk for privat deling, mens tilleggsinformasjon til tidsskriftsartikler, lab eller personlige nettsteder og fagspesifikke depoter eller dataarkiv blir brukt til å dele data offentlig (Allagnat et al., 2019).

En av utfordringene som dukker opp når det kommer til å dele data, er misbruk. Forskere finner det utfordrende å stole på andre. Tanker om misbruk av data eller om noen som velger å stjele data for å presentere det som eget er faktorer som gjør forskere tvilende til om de vil dele sine rådata eller ikke (Allagnat et al., 2019; Karasti et al., 2006). Selv om dette er en utfordring velger forskere fortsatt å dele data fordi de gjennom erfaringer har funnet ut hvor mye mer de kan vinne på å dele data istedenfor å holde de igjen (Karasti et al., 2006).

En annen utfordring enn ferdighetene til å dele data er det personlige aspektet med samarbeid. «Tilgang og deling av data er midlertidig også en kompleks sosial prosess der forskere må balansere ulike press og interesser (Karasti et al., 2006). Ulike press og interesser som påvirker forskerne vil variere fra prosjekt til prosjekt. På en side vil offentlige prosjekter der folket betaler for dataen gi et stort press for å dele data, men et forskningsprosjekt som er kjøpt av en leverandør for å vinne andeler eller patenter kan føre til en interesse om å holde igjen data. Videre vil erfaringer med å dele data og se fordelene i dette være med å redusere forskernes tilbakeholdenhet med å dele data fordi

de får en bredere forståelse av fordeler og ulemper ved å tilgjengeliggjøre de på nett (Karasti et al., 2006).

Videre mener forskere hvordan enkelte faktorer knyttet til datadelingspraksis er uviktig (Hrynaszkiewicz et al., 2021). Enkle praksiser som å lagre data i depoter eller databaser blir ansett som uviktig og derfor foreslår noen å bruke opplæring for å gi forskere ferdigheter og forståelse om viktigheten av dette. Ved å kunne integrere en slik løsning med den tradisjonelle publiseringen vil ikke nødvendigvis forskerne trenge å endre atferd når de publiserer. Majoriteten av forskerne blant PLOS-forfattere forteller hvordan tilleggsfiler til publikasjoner er hovedsakelig beskrivelser av hvordan de deler data, noe som kan gjenspeiles i hvordan 51% av forskerne i en studie deler offentlig data som supplement til deres journalartikler (Allagnat et al., 2019; Hrynaszkiewicz et al., 2021). Derfor kan en slik løsning hvor den tradisjonelle opplevelsen for dataløsning være optimal.

2.5 Byråkratiske regler om eierskap

Som nevnt tidligere har forskere utfordringer knyttet til eierskap og data og det kan være bekymringer for hvordan noen skal misbruke dataene (Allagnat et al., 2019; Karasti et al., 2006). Det eksisterer regler for hvem som eier data og hvem som kan bruke disse vederlagsfritt. DESCA er den mest utbredte modellen for Consortium Agreement for prosjekter som er en del av Horizon 2020 og er en modell som tar hensyn til krav som stammer fra reglene i det nye EU-rammeprogrammet Horizon Europe, samt erfaringer fra brukerne i arbeidet med DESCA (Helmholtz & Fraunhofer, 2022). Målet med modellen er å gi prosjektdeltakere sikkerhet selv om hvert prosjekt er forskjellig. Dette gjelder også sikkerhet når det kommer til eierskap av data.

Seksjonene som omhandler resultater, tilgangsrettigheter og ikke-avsløring av informasjon er seksjoner hvor avtaler om hvordan data kan bli brukt, hvem som kan bruke dataene og hvordan de kan brukes vederlagsfritt eller mot betaling kan avtales mellom partene.

2.6 Oppsummering

I kapittel 2 er databehandling beskrevet som en strømlinjeformet prosess og temaene er presentert som en strøm fordi det er gjort slik i forskningen. Det første steget for databehandling er datainnsamling. Datainnsamlingen har endret seg mye de siste årene og forskere har et bredt spekter av utvalg for hvordan de ønsker å samle inn data til sine prosjekter. Med de nye mulighetene har også rollene i et forskningsprosjekt endret seg. Dataforskere kan ta et større ansvar for datainnsamling, spesielt innen data som kommer fra forbrukere på internett, mens forskere i felt kan bruke mer avanserte sensorer og verktøy (misqresearchcurations, 2022).

På en annen side, skaper mulighetene spørsmål forskerne må besvare når det kommer til lagring og kvalitetssjekker. Normalt lagres data under lengre forskning i regneark eller lokale databaser (Diepenbroek et al., 2014). Selv om det eksisterer mange rammeverk for hvordan forskere kan lagre data i slike systemer, er det ofte for mange valgmuligheter og forskerne vet ikke hvordan de skal bruke dem på best måte (Alves et al., 2018).

Videre er det viktig å gjennomføre ordentlige kvalitetssjekker for å sikre god data. Det genereres utrolig mye data og forskere har mulighet til å bruke de. Utfordringen er hvordan de eksisterende kvalitetssjekksystemene ikke er kompatible med de vanlige

innsamlingssystemene (Cai & Zhu, 2015). Figur 2 viser til hvilke dimensjoner og elementer som er nødvendig for å kunne sikre god kvalitet i dataen. Disse dimensjonene og elementene vil være viktig å oppfylle for å sikre god kvalitet.

Avslutningsvis viser forskning hvilke utfordringer og fordeler forskere finner ved å dele data. Det eksisterer regler for hvem som eier dataen som deles, hvem som kan bruke de vederlagsfritt og det er muligheter for å inngå avtaler i prosjektene. Fordelene med å dele data er fremgang i forskningsmiljøet og muligheter for gjenbruk (Allagnat et al., 2019). Utfordringene til deling av data ligger i den komplekse sosiale prosessen hvor forskere føler dataen kan bli misbrukt (Allagnat et al., 2019; Karasti et al., 2006), uviktigheten av å lagre data i depoter (Hrynaszkiewicz et al., 2021) og utfordringer knyttet til hvordan dele data i praksis (Allagnat et al., 2019).

3 Casebeskrivelser

Det er utarbeidet to casebeskrivelser som beskriver situasjonen hos de to utforskede institusjonene. Forskningsmiljøet i det ene caset for økologisk forskning har søkelys på miljøforskning hovedsakelig innen luft, og gjennomfører majoriteten av sine prosjekter på bestilling av det offentlige. Forskningsmiljøet fra industriell forskning, har gitt tilgang til batteriforskning og forskning innen hydrogen. Majoriteten av prosjektene til caset som omhandler den industrielle forskningen er forskningsprosjekter i samarbeid med andre institusjoner, næringsliv eller universiteter.

3.1 Casebeskrivelse økologisk forskning

Forskerne i institusjonen fra dette fagområdet er veldig opptatt av databehandling og har en egen avdeling som jobber med å behandle data som blir tilgjengeliggjort i en åpen database. Forskningsprosjektet vil være med å forklare og beskrive utfordringer knyttet til aktivitetene rundt databehandlingen opp mot databasen, samt hvordan de jobber i felt og på lab. Institusjonen holder til inne på et avsperret området hvor bare de ansatte har tilgang. Hovedsakelig gjør de forskning innen klima og miljø for det offentlige, men noen av prosjektene er også finansiert privat gjennom bedrifter eller andre interessenter.

Denne institusjonen har gitt tilgang til et forskningsmiljø innen miljøovervåkningsdata og forskere som arbeider med å behandle data i en større database. Alle informantene har forskjellige utfordringer knyttet til databehandling, men det er eksisterer større, sammensatte utfordringer hvor det er problemer knyttet til data og hvordan denne lagres, bearbeides og kommuniseres til hverandre.

Majoriteten av dataen som blir hentet inn og behandlet blir gjennomført med offentlige midler eller på oppdrag fra det offentlige. Dermed må all data tilgjengeliggjøres for offentligheten. Fokuset på å dele kvalitetssikret data som kan brukes i rapporter settes til utregnede grenseverdier og det er både digitale og menneskelige kvalitetssjekker som er med å danne et grunnlag. Selv om forskningsinstitusjonen har digitale parametere og rutiner for å sikre kvalitet ønsker de å digitalisere flere prosesser og redusere bruk av forskjellig programvare. Endringene skal ikke føre til lavere kvalitet på data, men gi forskerne enklere forutsetninger for å finne variablene som påvirker dataene.

Rekrutteringen for caset var drevet av å finne informanter som ønsket å dele informasjon og tanker om hvordan de jobber med databehandling i sitt forskningsmiljø. Samtidig var det ønskelig å få informanter som ønsket å fortelle om sine vaner og hvordan de utarbeidet sine vaner gjennom læring og jobb. Videre var det viktig å finne informanter med forskjellig bakgrunn, stilling og fartstid i forskningsmiljøet for å få beskrivende variabler for oppgaven.

3.2 Casebeskrivelse Industri

I denne institusjonen beskrives databehandling som den store elefanten i rommet og dette forskningsprosjektet vil være med å besvare en liten del av den store utfordringen. Organisasjonen holder til i leide lokaler ved en utdanningsinstitusjon og en har en nyåpnet batteri-lab hvor de gjennomfører flere forskjellige prosjekter. Prosjektene kan

være alt fra å gjennomføre forskning for å kunne få patenter eller utarbeide forskningsartikler for å gi empiri eller teori som bidrag innen sitt forskningsmiljø.

Forskningsmiljøene organisasjonen har gitt tilgang til er innen batteri og hydrogen. Begge disse forskningsmiljøene har forskjellige utfordringer når det kommer til databehandling, men også flere likheter. Basert på den gitte definisjonen av databehandling jobber majoriteten av forskerne på de samme premissene med noen unntak.

Fordi organisasjonen jobber med andre forskningsinstitusjoner, universiteter og annen industri har de forskjellige utfordringer når de kommer til å dele data, hvilke byråkratiske retningslinjer de er pålagt å følge og hvem de faktisk ønsker å dele sine data med for å kunne sikre sine egne patenter. Derfor blir kvalitet, sikkerhet og deling av data et stort fokusområde for organisasjonen og de ønsker å finne ut av om de gjennomfører databehandling på en god måte og hvilke tiltak de eventuelt kan gjøre for å kunne skape bedre databehandling i sin organisasjon.

Rekrutteringen av informanter fra denne casen har vært drevet av å finne informanter som ønsker å dele informasjon om hvordan de jobber med data og hvorfor de har utarbeidet seg de forskjellige vanene og metodene de bruker. Et annet viktig aspekt har vært å finne informanter med forskjellige roller, erfaring og antall år i organisasjonen for å få kunne være beskrivende variabler til oppgaven.

4 Metode

Dette kapittelet beskriver forskningsmetodene som er benyttet under prosjektet. Først beskrives forskningsstrategien og hvorfor den tilnærmingen er valgt. Videre beskrives prosessen for hvordan informanter har blitt rekruttert og hvordan datainnsamling er gjennomført. Til slutt beskrives metodene for datanalyse, forskningsparadigme og metodevalget blir evaluert

4.1 Forskningsstrategi

Med bakgrunn i forskningsspørsmålene har dette forskningsprosjektet som mål å bidra med en empirisk innsikt og et forslag til endringer i praksis til hvordan forskere tilgjengeliggjør og beskriver data ved bruk av løsninger som tilrettelegger for samarbeid og samhandling. For å studere prosessen, metodene og føringene hos forskere som jobber med dette fenomenet er det valgt å gjennomføre to casestudier som belyser hvert sitt miljø og hvordan de jobber. I hvert av casene, er konteksten forskjellig og en multippel casestudie vil gi muligheten til å forstå likhetene og forskjellene mellom dem (Baxter & Jack, 2015). Deretter skal miljøene sammenlignes for å se etter likhetstrekk og mulighetene for å lære av hverandre. Den empiriske casestudien vil gi muligheten til å gå i dybden hos forskningsmiljøene og besøke forskere i deres naturlige setting der de jobber med data for å hente erfaringer og tanker.

Videre har den multippel casestudien åpnet muligheten til å møte mange informanter i forskjellige forskningsroller med forskjellig erfaring i sine miljøer for å danne et bredt spekter av tanker om fenomenet. Forskningsprosjektet går over et semester. Derfor er det valgt å se på hvordan forskerne jobber nå og finne forslag til hvordan de kan jobbe i fremtiden. Et alternativ til den valgte forskningsstrategien er bruk av spørreundersøkelse. Fordi det er viktig å finne tanker, erfaringer og holdninger til undersøkende fenomenet og undersøkelsers utfordringer med å gå i detalj er det valgt å gjennomføre casestudier istedenfor (Oates et al., 2022; Tjora, 2021).

4.2 Rekruttering av deltakere

Rekruttering av deltakere har blitt gjennomført på to forskjellige måter basert på de forskjellige casene.

For rekruttering til industriell forskning, har det vært to nøkkelpersoner som har hjulpet til å finne informanter. Den ene nøkkelpersonen ga tilgang til nøkkelperson to og inviterte til å presentere forskningsprosjektet på sitt gruppemøte. Etter presentasjon i gruppemøtet kom det flere informanter til som ønsket å være en del av forskningsprosjektet.

Nøkkelperson to i caset er selv en informant, men foreslo også flere informanter som kunne være relevante med tanke på å være beskrivende variabler. Ved å ha beskrivende variabler vil det være forskjellige erfaringer og stillinger som jobber med databehandling i organisasjonen. Disse har forskjellige erfaringer og vil kunne gi et godt perspektiv for å kunne si noe om vaner, metoder og rammeverk.

I forskningsmiljø økologi, ble det sendt ut e-post til forskjellige forskere basert på en liste med forskere som har vært samarbeidsvillige tidligere. Flere av forskerne på denne lista jobber ved institusjonen hvor det er dannet et case. Her er det opprettet en nøkkelperson med rollen som kontaktperson og hen hadde ansvaret for besøket under innsamlingen. Denne nøkkelpersonen har gitt tilgang til en rekke forskere som syntes prosjektet er interessant og derfor vil være en del av datainnsamlingen.

Før rekrutteringsprosessen kunne starte, måtte etiske hensyn med tanke på personvern bli adressert. Derfor ble det opprettet en prosess gjennom Norsk senter for forskningsdata slik at deltakerne kunne få en innsikt til hva de ble med på og hvilke rettigheter de har. Samtidig er alle intervjuer og observasjoner blitt anonymisert og lagret sikkert på NTNU sine servere.

Hovedsakelig er det forskere som har vært rekruttert fra de to forskjellige casene. Det finnes to unntak. Det første unntaket er en front end programvareutvikler som jobber med å lage data lesbart, og det andre unntaket er en leder som har ansvaret for digitaliseringen i sin forskningsavdeling

Rekrutteringen har vært drevet av et ønske om å få nok data fra forskjellige forskere og informanter i relevante stillinger med forskjellige perspektiver på databehandling for å kunne gjennomføre forskningsprosjektet på best mulig måte.

4.3 Datainnsamling

For datainnsamlingen er det valgt å gjennomføre observasjoner av forskere med muligheter for å holde intervjuer i etterkant og stille spørsmål i form av ustrukturerte intervjuer under observasjonen. Det er valgt å bruke disse to kvalitative tilnærmingene for å finne svar på forskernes holdninger knyttet til fenomenet forskningsprosjektet prøver å besvare. Observasjonene skal besvare hvordan forskerne arbeider med databehandling og intervjuene skal hjelpe til å forstå hvorfor de arbeider på den måten. Som supplement til de to hovedmetodene, er uformelle samtaler, observasjon av møte og besøk på lab gjennomført for å få en større og generell innsikt i hvordan forskere jobber arbeider med databehandling.

Type observasjon	**Case økologisk forskning**	**Case industriell forskning**	**Total**
Observasjon med ustrukturert intervju (45-90 minutter pr sesjon)	0	6	6
Ustrukturert intervju (ca. 1 time pr intervju)	6	1	7
Besøk på lab (ca. 45 min pr lab)	2	2	4
Uformell samtale (ca. 45 min pr samtale)	1	3	4
Observasjon av møte (ca. 1 time)	1	0	1
Sum	**10**	**12**	**22**

Tabell 1 En oversikt over datainnsamling.

Tabellen viser til hvor mange datainnsamlingssesjoner som er gjennomført med de individuelle casene og totalsummen av casene. Det er viktig å påpeke tabellen viser til sesjoner med datainnsamling og ikke antall informanter. Noen informanter har deltatt i flere av sesjonene. Et eksempel på dette er nøkkelpersonene som er nevnt under rekrutteringen. I økologisk forskning har nøkkelpersonen vært med deltakende informant i observasjon av møte, uformell samtale og ustrukturert intervju. Fra industriell forskning har en av nøkkelpersonene vært med som informant under uformell samtale, ustrukturert intervju og besøk på lab, samt den andre nøkkelpersonen var med på en uformell samtale og en observasjon med ustrukturerte intervjuer. Hovedsakelig er det disse tre nøkkelpersonene som går igjen, men fra økologisk forskning er det til sammen tre stykker som har vært deltakende informanter under observasjon av møte og deltatt i ustrukturerte intervjuer.

For å opprettholde anonymiteten til informantene, er de gitt et tilfeldig nummer i innenfor sitt case. Dette er fordi informantene kjenner hverandre (fordi de er kollegaer), og kan ha hatt samtaler seg imellom om når de var en del av datainnsamlingen. Med bakgrunn i dette ble det besluttet å ikke, nødvendigvis, bruke «Informant 1» om den første informanten som deltok i prosjektet, men heller endre tallene til informantene gjennom en random number generator.

4.3.1 Hovedmetoder for datainnsamling

Den første hovedmetoden for datainnsamling skal svare på hvordan informantene arbeider. For å svare på denne delen av fenomenet er det valgt observasjon. Observasjonen som er gjennomført er en form for dynamisk observasjon der informantene har vist og beskrevet hvordan de gjennomfører sine arbeidsoppgaver. Denne formen for observasjon beskrives av Oates et al., (2022) som en «participant observation» der forskeren tar en del av fenomenet som skal undersøkes for å få erfaringer fra informantene sitt ståsted (Oates et al., 2022). I prosessen er det satt av tid til å kunne lære mest mulig. Samtidig er det umulig å huske alt som skjer og derfor er det skrevet ned notater under observasjonene ved tilfeller det kommer opp noe interessant samt noen tanker rundt det som kom opp.

Basert på de fire observasjonsmetodene som Oates et al. nevner er det ingen som kan relatere til hvordan denne observasjonen er gjennomført. Den nærmeste metoden vil være en komplett observatør hvor forskeren følger informanten og observerer alt som skjer, men ikke er delaktig i prosessen (Oates et al., 2022). Denne tilnærmingen er den som er brukt, men informantene er stilt spesifikke spørsmål på hvordan de gjennomfører spesifikke oppgaver. Selv om informantene er pekt i en retning har de blitt komplett observert og det er informantene selv som har gjennomført og beskrevet sine arbeidsoppgaver relatert til databehandling.

Sammen med observasjoner er det valgt å bruke semi-strukturerte intervju for å underbygge det som kommer frem. Dermed skal denne formen for datainnsamling være med å svare på hvorfor informantene jobber på måten de gjør, og hvilke holdninger de har til arbeidsmetodene sine. For de semi-strukturerte intervjuene er det utarbeidet en intervjuguide som grunnlinje for å finne svar på hvilke holdninger og tanker informanten har rundt sine arbeidsoppgaver. Det sentrale med det semi-strukturerte intervjuet er å innhente skildringer av livsverden til den intervjuede og hvordan informanten opplever fenomenet (Krumsvik 2015). Fordi det er ønskelig å innhente beskrivelser av denne opplevelsen, vil et semi-strukturert intervju der det er lov å gå litt utenfor de satte rammene være fordelaktig for å besvare dette.

I tillegg til de to hovedmetodene er det gjennomført en dokumentgjennomgang. Under observasjonene er det gitt tilgang til dokumenter for å få innsikt i testprosedyrer og databehandling relatert til testprosedyrene. Det er også gjennomgått byråkratiske dokumenter som beskriver hvordan en avtale om eierskap til data kan se ut.

4.3.2 Andre metoder for datainnsamling

Som supplement til observasjonene og de ustrukturerte intervjuene er det gjennomført andre datainnsamlinger med andre metoder. Metodene er brukt til å få en forståelse av hvordan forskjellige instrumenter fungerer, og hvilke utfordringer forskere står ovenfor i fasene før dataen er samlet inn. Samtidig har metodene gitt innsikt i infrastrukturen på arbeidsplassen til forskerne, hvilke utfordringer de tar opp med hverandre på gruppemøte og hvilke utfordringer de føler forskningsprosjektet kan være med å belyse.

Metodene for å innhente denne informasjonen er å observere et gruppemøte, besøke labber og ha uformelle samtaler med forskere om prosjektet. Observasjonen på gruppemøte var en komplett observasjon (Oates et al., 2022), der informantene diskuterte utfordringer knyttet til innføring av et nytt system.

4.3.3 Gjennomføring av datainnsamling

Gjennomføringen av datainnsamling startet med uformelle samtaler. Samtalene var med å gi en bedre innsikt i utfordringer knyttet til databehandling. Spesielt organisasjonen i caset for industriell forskning, var veldig åpne i denne fasen om hvilke utfordringer de følte på og hvordan forskningsprosjektet kunne være med å belyse og være til hjelp for deres utfordringer. Som en del av samtalene ble det første besøket på lab gjennomført og det ble diskutert infrastrukturelle utfordringer. Basert på samtalene ble det utarbeidet en intervjuguide som ble brukt for å finne svar på holdninger og erfaringer som beskriver fenomenet prosjektet prøver å belyse og besvare.

Deretter ble det gjennomført observasjoner og ustrukturerte intervjuer med industriell forskning for å hente data om fenomenet. Observasjonene og de ustrukturerte intervjuene ble gjennomført på informantene sin arbeidsplass. Informantene fikk selv velge hvilke oppgaver de ville gjennomføre og hvor på arbeidsplassen de ville gjennomføre oppgavene. Dette er spesielt viktig når det kommer til intervjuer fordi informanten burde være komfortabel i settingen for å kunne dele mest mulig av sine erfaringer om fenomenet (Tjora, 2021).

Videre ga økologisk forskning tilgang til sine forskere, men det var ikke mulig å gjennomføre en til en observasjoner som i industriell forskning. Derfor ble intervjuguiden endret for å legge til rette for en datainnsamling der det i tillegg til søkelys på hvilke metoder som ble brukt og hvorfor de metodene ble brukt. De ustrukturerte intervjuene ble gjennomført på et møterom hos institusjonen til informantene igjen for å skape en komfortabel setting (Tjora, 2021). Under de ustrukturerte intervjuene i økologisk forskning, spurte majoriteten av informantene om hvordan gjennomføringen av intervjuet ville være. Derfor ble det gjennomgått deres rettigheter med tanke på NSD og hvordan observasjonene med ustrukturerte intervjuer i industriell forskning var gjennomført. Deretter ble de forklart hvordan intervjuguiden var et hjelpemiddel for å besvare de samme spørsmålene som observasjon sammen med intervju svarte på i den andre casevirksomheten. Økologisk forskning ga også tilgang til to av sine labber under besøket for å finne ut av hvordan instrumenter fungerte og hvordan infrastrukturen så ut.

Til slutt er det gjennomført et nytt besøk hos industriell forskning. Besøket var på den nyeste laben og det var mulig å få et innblikk i hvordan de nyeste instrumentene og maskinene er plassert og hvordan tilgangen vil være samt hvilke utfordringer som kan oppstå.

4.3.4 Opplevelse av datainnsamling

Datainnsamlingen har vært utfordrende av flere årsaker. I starten var det vanskelig å balansere observasjon med spørsmål. En av nøkkelpersonene var første informant til observasjon og biveileder var med på første sesjon for å være med å skape en trygg ramme for økten. Det ga stor læring og dermed var det mulig å gjennomføre datainnsamlingssesjoner alene i ettertid fordi det var arbeidet inn større erfaring med datainnsamling.

En annen utfordring knyttet til opplevelsen av datainnsamling er personlig forskjeller. Som mennesker er også forskere forskjellige. Dette gjenspeiler seg i hvor pratsomme de enkelte informantene er. Noen informanter kan være veldig åpne og innbydende mens noen er vanskeligere å få informasjon ut av. Konsekvensen av forskjellene kommer frem i hvor mye relevant data de forskjellige informantene får delt. Hos de som prater mest er det kommet veldig mye data som beskriver fenomenet på en god måte, men også mye data som det er vanskelig å plassere opp mot forskningsspørsmålene.

På en annen side har informantene som prater mindre svart eksakt på spørsmålene og delt lite av erfaringer og holdninger. De nevnte informantene virket ukomfortable i situasjonen og det var derfor vanskelig å vite hvor dypt det var mulig å dykke for å hente ut informasjon fra dem.

Den siste utfordringen som oppstod var tilgang på båndopptaker. I de tre observasjonene som er gjennomført sammen med biveileder, er det gjennomført lydopptak i tillegg til tatt notater. I de resterende datainnsamlingssesjonene er alle notatene tatt for hånd uten båndopptaker noe som kan ha ført til tap av data.

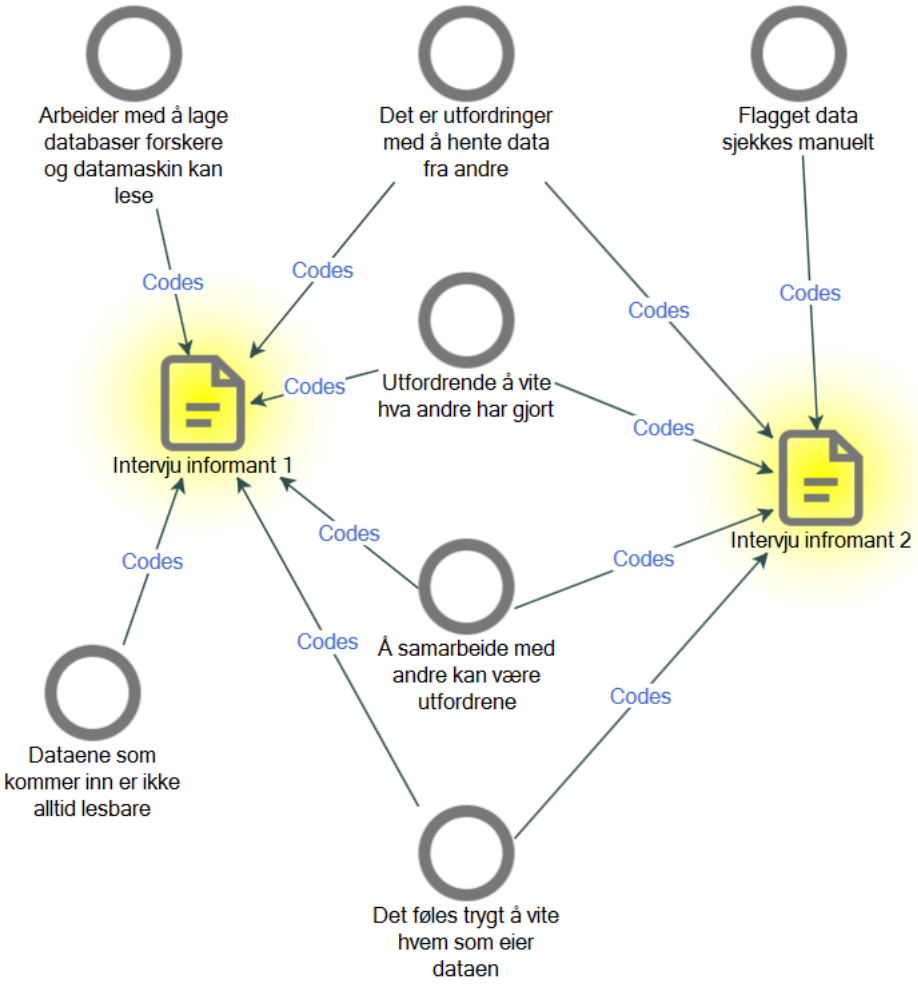
4.4 Metode for dataanalyse

For å analysere dataene som er innhentet, er det valgt å kode svarene fra informantene. Kodingen er gjennomført ved bruk av NVivo og det er brukt en guide som er vedlagt boken Kvalitative forskningsmetoder i praksis (Tjora, 2021). Guiden starter på side 321, og forklarer i steg for steg hvordan programvaren skal brukes. Det er denne steg for steg bruksanvisningen som er brukt selv om det er en guide for NVivo 1.3 og versjonen som er brukt for å kode dataene i dette forskningsprosjektet er NVivo 2.7.1. Videre er det valgt å bruke SDI-modellen fordi det vil holde fast ved en ren induktiv strategi, samtidig som det vil hente ut de mest sentrale dataene, redusere volumet av data og legge til rette for et idegrunnlag (Tjora, 2021).

Det er valgt å bruke en empirinær kodings strategi for å kunne beskrive hva informantene følte istedenfor å si hva de pratet om. Dermed vil det være mulig å bruke empirikodene alene for å beskrive interessante funn til analyse og diskusjon (Tjora, 2021). En annen fordel som blir beskrevet ved å bruke denne tilnærmingen, er hvordan svarene vil være mulige å finne i kodene istedenfor å gå tilbake til transkripsjonene av intervjuene og observasjonsstudiene for å hente informasjon (Tjora, 2021).

For å sikre god empirisk koding er det gjennomført en kodetest. Kodetesten består av to spørsmål der begge må bestås for å kunne antyde god induktiv SDI-koding (Tjora, 2021).

> «Spørsmål 1: Kunne man laget koden *før* kodingen?
> a: hvis ja: a priori (unødig) koding – lag annen kode!
> b: hvis nei: potensielt god empirinær koding – gå videre til spørsmål 2!
>
> Spørsmål 2: Hva forteller *bare* koden?
> a: tematiser datasegmentet (fra intervju: hva det ble snakket om): unødvendig sorteringskoding – lag annen kode!
> b: gjenspeiler konkret innhold (fra intervju: hva som ble sagt): god empirinær koding! (Tjora, 2021)»



Figur 3: Utklipp av kodesammenligningen fra NVivo. Sirklene viser til kodenavn og de rektangulære viser til hvilken informant. Pilene viser sammenhengen mellom kode og informant.

Figur 3 er et eksempel på hvordan to informanter kan sammenlignes i NVivio. I utklippet er det valgt å bruke to informanter og 7 koder for å gi innblikk i hvordan sammenligningen og den empiriske kodingen ser ut, istedenfor å vise sammenligningen av alle kodene og de 22 datainnsamlingssesjonene. Kodene er tilknyttet en eller flere informanter avhengig av utsagn eller observasjon. Videre er de testet opp mot

spørsmålene fra Tjora sin kodetest og på bakgrunn av alternativ b i begge spørsmålene er det valgt å gå videre med kodene til kodegruppering med god empirinær koding.

Det neste steget var gruppering av kodene for å danne konsepter som vil svarer på fenomenet. Kodegrupperingen er gjennomført induktivt, og har bestått av å samle koder som er tematisk like og gruppere de før unødvendige koder er etterlatt i en restgruppe (Tjora, 2021). Med bakgrunn i denne kodegrupperingen er det skapt konsepter som er med å besvare informantenes holdninger og tanker om fenomenet.

4.5 Forskningsparadigme

Forskningsprosjektet er en kvalitativ flercasestudie med mål om å tilføre empiriske data og eventuelt foreslå deler av et rammeverk for hvordan forskere arbeider med databehandling. Oppgaven har fått et sosio-teknisk perspektiv fordi hver av organisasjonene i casene ansees som totalsystemer bestående av ett sosialt- og ett teknisk system (Sander, 2021) for å forstå og beskrive relasjonen mellom teknologi og mennesker. Det er valgt å bruke en interpretivistisk tilnærming i oppgaven for å bedre forstå hvordan det sosiale og teknologiske aspektet spiller sammen. Forskningen er dermed opptatt av å forstå og beskrive denne sosiale konteksten til teknologien (Oates et al., 2022).

Videre er mulighetene for å tolke dataene på forskjellige måter og refleksjon over forskerrollen begrunnelser som tilser hvorfor en interpretivistisk tilnærming er riktig (Oates et al., 2022). Dataene fra intervjuene vil være mulig å tolke forskjellig basert på forskerens bakgrunn. Samtidig kan forskeren ha en påvirkning på informantene og deres holdninger under intervjuer og observasjon, selv om det er jobbet for å innta en passiv rolle i datainnsamlingssituasjonen. På en annen side, er tema og spørsmål under observasjon og intervju styrt etter utarbeidet intervjuguide med muligheter for å stille oppfølgende spørsmål under datainnsamlingen (Oates et al., 2022; Tjora, 2021).

4.6 Evaluering av metodevalg

Metodens validitet kan ses i sammenheng med metodens relabilitet. I kvalitativ forskning, vil dette bety om det ønskede utforskede fenomenet faktisk er forsket på (Krumsvik, 2015). Krumsvik presenterer sju forskjellige, men tilpassede faser som skal være med å gi forskeren et grunnlag for å validere sitt kvalitative verk.

Prosessteoretisk forankring: Fasen går ut på å beskrive teoretisk robusthet i oppgaven. Videre sier validiteten noe som sammenhengen mellom forskningsspørsmålene og det presenterte teorigrunnlaget.

Planlegging: Fasen omhandler valg av forskningsmetodikk for å finne svar på forskningsfenomenet om gjennomføringen er tilfredsstillende.

Intervjuing: Fasen omhandler samspillet mellom forsker og intervjuobjekt. Validiteten vil dermed si noe om troverdighet, utarbeidelse av intervjuguide og kvalitetssjekk.

Transkribering: Fasen handler om hvor ekstremt viktig det er å gjengi intervjuobjektene eksakt.

Analyse: Fasen omhandler om det er gode spørsmål i intervjuene og om svarene tolkes på en god måte.

Validering: Fasen omhandler om det er reflektert riktig i forhold til hvilke former for validering som faktisk er relevante for studien og hvilke tiltak som er gjort for å sikre validering.

Rapportering: Fasen omhandler om rapporten gir en god skildring av studien (Krumsvik, 2015).

4.6.1 Egenvurdering av metodevalg

Når det kommer til prosessteoretisk forankring er det sammenheng mellom forskningsspørsmål eller forskningsfenomenet og teorien som blir presentert. Den presenterte teorien skaper en god forståelse for hvordan databehandlingsprosessen blir gjennomført og hvilke utfordringer som eksisterer knyttet til databehandling. Det teoretiske grunnlaget gir videre god innsikt i teorier som drøftes opp mot resultatene i diskusjonen for å kunne besvare på forskningsspørsmålene.

Tidligere er det argumentert hvorfor det er valgt en kvalitativ tilnærming for prosjektet istedenfor en alternativ tilnærming med spørreundersøkelser. For å oppsummere er begrunnelsen for å bruke en kvalitativ tilnærming mulighetene til å finne tanker, erfaringer og holdninger til fenomenet. Med bakgrunn i mulighetene kvalitativ forskning gir for å besvare dette er det den valgte metoden i form av casestudier (Oates et al., 2022; Tjora, 2021).

Neste fasene er intervju og transkribering. Det er gjennomført en pilotstudie etter utarbeidet intervjuguide for å teste spørsmålene og eventuelt gjøre endringer. Det ble valgt å beholde intervjuguiden fordi den ble erfart som god og spørsmålene ga pilotinformantene mulighet til å prate fritt om temaene som ble presentert. Guiden ble endret mellom de to casene fordi det ene caset hadde tilgang til observasjoner under intervjuet. For å utjevne denne forskjellen ble det lagt til noen spørsmål om erfaringer rundt systemer.

Selv om intervjuguiden føles til å være av god kvalitet er det noen utfordringer knyttet til transkripsjon. I utgangspunktet er det ønskelig at resultatet er likt det intervjuobjektet mente, men det er to årsaker til hvorfor dette kan være litt unøyaktig i noen settinger. Den første årsaken er om informasjon er tapt eller endret under oversetting. Noen av datainnsamlingene er gjennomført på norsk, andre er gjennomført på engelsk. Under transkripsjonen kan dermed noe ha blitt tapt eller mistolket. Samtidig er noen av datainnsamlingssesjonene gjennomført med transkripsjon for hånd. Dermed kan små ord ha blitt borte, noe som mulig påvirker utsagnene i en liten grad. For å løse disse utfordringene er de eventuelle informantene kontaktet om det er en kontekst det er ønskelig å få svar på.

Dataanalysen er gjennomført med en kvalitativ analysemetodikk med SDI-metoden. Med denne metoden er det jobbet stegvis i etapper fra rådata til konsepter. Selv om modellen kan fremstå lineær, noe som forskningsprosjektet ikke har vært, har den vært med å danne et godt utgangspunkt for systematikk og fremdrift (Tjora, 2021). Ved bruk av denne modellen og teori fra Tjora er det gjennomført en kvalitativ analysemetodikk hvor rådataene fra datainnsamlingen er redusert til konsepter som er interessante og relevante for problemstillingen.

Videre er de fasene som er trukket frem relevante for oppgaven. Det er nevnt samspillet mellom teori og forskningsfenomen, hvorfor det er valgt kvalitativ forskning, hvilke

utfordringer som er møtt med transkripsjon og hvordan dataanalysen er gjennomført. Dette er interessante temaer som vil beskrive hvor valid en oppgave er.

Til slutt er det rapporteringsfasen. Rapporten er skrevet på en måte hvor leseren skal få mest mulig innsikt i hvordan prosjektet er gjennomført og hvorfor de forskjellige valgene som er gjort er besluttet. Det er lagt innsats i å skildre studien på best mulig måte for å skape en transparent studie hvor det er mulig å gjennomføre studien igjen basert på informasjonen som er vedlagt.

For å oppsummere virker rapporten å være et valid prosjekt basert på Krumsvik sine kriterier. Prosjektet er transparent, og valgene er begrunnet i teori. På en annen side er det utfordringer til deler av rapporten, men det er gjort tiltak for å sikre en datainnsamling som ikke bærer preg av disse.

5 Funn

Dette kapittelet beskriver funnene fra flercasestudiet. Resultatene blir presentert basert på casene. Videre er resultatene tematisk inndelt og inkluderer eksempler og/eller utdrag fra datainnsamlingen.

5.1 Case økologisk forskning

Delkapittelet beskriver hvilke resultater som er kommet frem i dette caset. Det er valgt å dele underkapitlene opp etter infrastruktur, opplæring og databehandlingsprosessen for å best diskutere og besvare forskningsspørsmålene.

5.1.1 Infrastruktur

Bygget hvor forskningsstasjonen holder til er godt sikret. For å komme inn, må besøkende vente på forskeren som har ansvaret for besøk og ekstern vakttjeneste slipper besøkende inn gjennom en egen dør. Samtidig er det flere porter som må åpnes på veien til bygget. Mellom to av portene ble det fortalt: «*Vanligvis er det færre sikkerhetstiltak, men på grunn av [utelatt] er det litt ekstra sikkert nå*» Informant 9.

Selve bygget er delt inn etter avdelinger. På den ene siden av bygget er forskningslabene og de som arbeider med lab sitter på den siden, mens dataingeniørene som jobber med databaser og datarensking sitter på den andre siden, fysisk avskilt av en trapp. Det er hovedsakelig to labber hvor det gjennomføres forskjellige typer arbeid basert på den økologiske forskningen. Den ene laben ble hovedsakelig brukt til luftforskning. Forskningsstasjonen har egne testprodukter som sendes til en caretaker i felt. Caretakeren trenger ikke å være en forsker, men har ansvaret for å utplassere testproduktet og sende testene tilbake til forskningsstasjonen. Deretter blir prøvene fra testproduktet hentet ut, testet og resultatene blir omgjort til data.

Under testingen sitter forskeren ved maskinen, overvåker dataskjermen eller gjør endringer på denne.

«*Jeg gjennomfører testen på stasjonen. Når testen er ferdig, sjekkes de opp mot modeller i et unikt nummer og sendes inn i databasen*» Informant 11

På den andre laben ble det arbeidet med andre former for målinger. Disse målingene blir gjennomført med andre systemer og kobles opp til en databoks. Databoksen og er egenprodusert og hele institusjonen bruker samme verktøy. Under omvisningen blir det fortalt hvordan Miljødirektoratet utvikler sin egen løsning, noe guiden mente var unødvendig fordi det allerede eksisterte et utviklet system. Utfordringen systemet er med å løse er hvordan det gjør de binære tallene, som er output, til lesbare data. Om en forsker ønsker tilgang til denne dataen må hen vente til dataen er evaluert før tilgang gis. En evaluering gjennomføres en gang i måneden.

5.1.2 Opplæring

Hvilken opplæring informantene har varierer. Noen har vært med på få kurs, andre har måttet lære seg selv eller spørre andre om hjelp. Det kommer frem i resultatene at forskningsinstitusjonen har lite fokus på etterutdanning eller annen opplæring utenom tradisjonell innføring i de forskjellige systemene som brukes.

Da informantene ble spurt om hvilken opplæring de hadde fått utenom sin utdanning som kunne hjelpe dem, ble det fortalt hvordan det var lite opplæring og hvordan opplæringen som ble gitt var lite strømlinje formet og veldig varierende

«Jeg er ikke kurset i institusjonen, men har blitt opplært av enkeltpersoner. Jeg får veldig mye forskjellige svar avhengig av hvem jeg spør, og det er ikke noe strømlinjet form for arbeidsvaner eller arbeidspreferanser» Informant 10.

«Jeg har vært på to kurs, men lærer mest i prosjekter og det er ikke ofte det har blitt tilbudt kurs.» Informant 13

«Jeg fikk en introduksjon til systemene vi bruker, men det har ikke vært noen kurs eller annen opplæring for jobben. Mye har vært opp til for meg selv å finne egne rutiner og måter å jobbe på for å få jobben gjort» Informant 8.

På grunn av organisasjonen sin tilnærming til opplæring var arbeidsmetoder svært varierende. De ansatte har sin egen metode for å gjennomføre arbeidsoppgaver og det er ingen overordnet veiviser som beskriver de beste praksisene på arbeidsplassen. Konsekvensen av tilnærmingen er hvordan ansatte med like arbeidsoppgaver gjennomfører sine oppgaver med varierende grad av suksess og de får forskjellige tilbakemeldinger på hvordan oppgaven skal gjøres ved å spørre forskjellige ressurspersoner.

5.1.3 Databehandlingsprosessen

Resultatene viser hvordan forskerne i organisasjonen har ansvaret for dataanalysen og behandling av data hvor det kan ha skjedd hendelser før de sender den inn til en større database. Når dataen er i databasen er det datamanagerne som har ansvaret for kvalitetssjekk og tilgangsstyring. Forskere som ikke er en del av forskningsinstitusjonen, men jobber på prosjekter som har avtaler eller som selv har en avtale kan også bruke databasen.

«Dataene vi henter inn er veldig sensitive. Om en person ute på Svalbard puster, vil det slå ut på testene. Derfor må de logge hver gang de går ut slik at jeg kan flagge dette i testdataene i etterkant. Det kan også oppstå hendelser som kan gi utslag på naturgasser i luften. Et ekstremt eksempel på dette var eksplosjonen på gassrørledningen hvor store mengder naturgass ble oppdaget. Hendelsen ga ekstreme verdier som slo ut på grenseverdiene.» Informant 12.

«Dataene som kommer inn til databasen gjennomgår en test hvor verdier som er over eller under grenseverdiene flagges. Det er disse verdiene vi går inn og sjekker. Om forskerne har beskrevet hvorfor kan det gå videre hvis ikke må vi kontakte forsker og finne ut av hvorfor de dataene krysser grenseverdiene. Etter dataene er godkjent ved skript eller manuelt eller begge får de en ID og en status. Før ble dette gjort på papir, nå er det ved databaser. Dette systemet gjør deling mer strømlinjet, unikt og effektivt» Informant 11.

Det er tydelig hvordan det er en fordeling mellom forsker og datamanager. Forskerne har ansvaret for data frem til den er i databasen mens de ansatte datamanagerne har ansvaret for de dataene som er i databasen. Dette fører til en dobbelt runde med kvalitetssjekk hvor forskerne først sjekker sine data før de blir sjekket av et script opp mot grenseverdier i det de blir sendt inn i databasen.

Selv med denne tilnærmingen er det utfordringer knyttet til innsending og script. Mye av dataene som kommer inn kan være i filformater eller det de kan som nevnt ha problemer med grenseverdier. For å løse dette jobbet datamanagerne i en problemløser hvor de kan kommunisere med forskeren som har sendt inn data. Problemløsningssystemet integrerer egne systemer med databasen og gjør det mulig å flagge data eventuelt legge til metadata som er med å beskrive hva som er sendt inn.

Mye av datadelingen hos forskerne skjer gjennom databasen og tilgang til databasen, men de ansatte i må også dele data og dokumenter med hverandre. Noen i institusjonen har en rolle hvor de er 50% forsker og 50% datamanager eller dataingeniør og jobber med utvikling og drift av databasen.

«Mye av kommunikasjon og samarbeid gjennomføres i Teams. Noen av bruken føles rotete når det kommer til prosjekter og da liker jeg … bedre. Det har vært en overgang med tanke på bruk og endelige resultater legges på OneDrive i form av rapporter mens data puttes i databasen. Verktøyene er hjelpsomme og Teams funker best til å brainstorming, dele dokumenter og oppsøke hjelp. Typisk henter jeg data fra databasen, plotter den i Python og deler resultatene i Teams i et dokument hvor vi samskriver.» Informant 8.

«Vi bruker Teams mye til chat og fordeling av oppgaver og dokumenter til møtevirksomhet. Det er også fint å kunne bruke Trelloboards for å se oppgaver og flytte rundt på dem ettersom de er ferdige eller ikke. På en annen side er de store kanalene som å rope i naturen. Ingen eierskap, det blir brukt som en stor oppslagstavle. Mindre grupper fungerer bedre.» Informant 10.

«Microsoft Teams blir mye brukt til møter, mail, kalender, OneDrive, SharePoint for filer og chatsamtaler om det skulle være utfordringer. Problemløsningssystemet brukes også for å samarbeide med forskerne for å løse utfordringer knyttet til dataen som blir sendt inn» Informant 9.

Resultatene viste varierende holdninger og bruk av samhandlingssystemer. De informantene som hovedsakelig brukte Microsoft Teams til kommunikasjon og enkel dokumentdeling følte løsningen ble brukt på en effektiv måte og det var enkelt å navigere seg. Andre mener det kan være vanskelig å navigere og at løsningen ikke funker til prosjekter større prosjekter og heller burde være for mindre grupper.

Videre kommer det frem i resultatene hvordan enkelte ønsker å digitalisere prosessene enda mer. Selv om dataene har gått fra papir til database og det er script som hjelper å finne grenseverdier, er det fortsatt mer som kan gjøres. Mer digitalisering vil føre til høyere kvalitet og mer strømlinjeformet arbeid.

Da informantene ble spurt om deling av data i sitt arbeid. Fortalte de om hvordan regler bestemte hvem som eier dataene de produserer og hvordan dette føltes trygt. Samtidig beskrev de hvordan forskningen i deres felt tidlig var ute og ønsket å dele sine funn med andre. Det ble også nevnt hvordan noen ønsker å drive åpen kildekode prosjekter med land som ikke har ressurser til å samarbeide for å kunne gi de nye impulser og muligheter.

«Denne typen forskning er tidlig ute med å dele data og grunntanken er at offentlig betalt forskning betyr offentlig tilgang til data. CB14 begrensninger ligger til hvor dataen er hentet fra og skal det skrives må det kontaktes. Det publiseres mye drit,

men vi eier ikke dataen i vår rolle. Er det oppdrag for industri er det oppdragiver som bestemmer, men hovedsakelig har vi offentlige oppdrag.» Informant 8

«Det vi deler skal være forutsigbart og så lite feilfritt som mulig. Jo mindre menneskelige feil, jo sikrere er vi dataen vi deler er bra.» Informant 10

«Å dele data med andre er det som vil drive forskningen fremover og jeg er stolt over dataene mine og om noen ønsker å bruke den. Samtidig bruker jeg data fra andre steder for å sammenligne og se hva vi kan lære av andre. Jeg stoler på at dataen jeg får tak i, gjennom databasen, er av kvalitet fordi vi har prosedyrer for alt og bruker standarder slik at alt nå gjøres likt.» Informant 12

Det er tydelig hvordan majoriteten informantene ønsker data av kvalitet og føler seg trygge på at forskningsinstitusjonen gjør en god jobb for å sikre dette selv om de kan bli bedre. Samtidig kommer det frem hvordan de hele tiden ønsker å forbedre seg for å gjøre fremskritt innen forskningsområdet ved å bruke egne og andres erfaringer for å hente og dele data og datasystemer.

For å oppsummere databehandlingsprosessen er denne todelt. Forskerne har ansvaret for datainnsamlingen, og den første kvalitetssjekken før datamanagerne overtar ansvaret i databasen. Før dataen blir godkjent i databasen gjennomgår den sjekker opp mot grenseverdier. Om den overgår grenseverdiene blir dataen flagget og forskeren må forklare hvorfor. Deling av data skjer gjennom tilgang til databasen og fordi majoriteten av prosjektene er offentlige vil alle ha tilgang.

5.2 Case industriell forskning

Delkapittelet beskriver hvilke resultater som er kommet frem i dette caset. Det er valgt å dele underkapitlene etter samme inndeling som 5.1 for å skape et uniformt resultatkapittel som er med å danne grunnlaget for diskusjon av forskningsspørsmål.

5.2.1 Infrastruktur

Denne forskningsinstitusjonen brukte to labber under datainnsamlingen. Den ene laben var en del av en utdanningsinstitusjon og derfor hadde ikke forskningsinstitusjonen helt kontroll på tilgangsstyringen. Videre eide ikke institusjonen kontroll over nettverk og var avhengig av utdanningsinstitusjonen sitt nettverk for tilgang til internett. Dette var en utfordring fordi utdanninginstitusjonen gjorde oppdateringer på nettverket eller tok det ned på nattestid for å gjøre oppdateringer. Da testmaskinene ikke var koblet til nettverket, ble testprogrammene stoppet. Den gjentatte hendelsen er en utfordring som skaper merarbeid for forskerne. Som en løsning på utfordringen, kan testene kjøres uten å være på nettverk fra starten av. Dermed kan institusjonen gjennomføre lengre tester om de ønsker.

«En test som varer i 500 timer og kanskje nesten var ferdig må gjennomføres på nytt» Informant 7.

På en annen side, skaper løsningen en ny utfordring. Selv om forskeren kan se på dataen manuelt på teststasjonen får de ikke hentet dataene over på sin egen datamaskin. Om en forsker ønsket å hente data fra testmaskinen over til egen datamaskin brukte de en gratis og ikke-kommersiell programvare for fjerntilgang. Denne løsningen lot forskeren

hente data fra testmaskinen over til sin egen datamaskin eller overvåke testmaskinen fra kontoret og/eller hjemmekontor. Problemet med løsningen var nødvendigheten av internett for å tilgang. Dermed går forskningsinstitusjonen i loop med hvilken utfordring de ønsker å stå ovenfor. Informant 3, 2, 1 og 5 forteller hvordan de finner det utfordrende å hente data fra testmaskiner som ikke var på internett fordi det ikke var noen utbredt policy i institusjonen om hvordan det skulle gjøres.

Den andre forskningslaben hadde ikke denne infrastrukturelle utfordringen fordi det er ikke en annen institusjon som eier området på samme måten.

«Målet med denne laben er å kunne samle all data i en fil og poenget med systemet å gjøre data mer uniforme. Vi ser på flere leverandører og software for å sammenligne for å finne ut av hvordan det kan bli bedre. Fordelen er å få med prosessdataen i den nye laben» Informant 7.

Den nye laben var ikke helt ferdig under datainnsamlingstidspunktet. Selv om utsagnet viser til målene for laben og hva den kan tilby var ingen store avgjørelser tatt da det kom til system. For informanten var det også vanskelig å svare på om de ønsket å bruke samme programvare for fjerntilgang og overvåkning eller om de ønsket å bruke en annen løsning.

5.2.2 Opplæring

Resultatene viser hvordan forskningsinstitusjonen har lav innsats innen opplæring. Forskerne blir gitt en introduksjon til hvordan noen av programvarene fungerer og en innføring i hvordan de skal bruke de forskjellige testmaskinene. Det blir også gitt en innføring i hvordan de kan notere og protokollføre viktig informasjon om testene.

Da informantene ble spurt om hvilken opplæring de fikk utenom utdanning, ble det beskrevet som enkle innføringer i bruk og protokollføring. Det ble også fortalt hvordan hver de har prøvd mye selv og hvem som ga de innføringen har påvirket deres arbeidsmetoder i stor grad.

«Jeg fikk en innføring av min forskermentor i hvordan bruke testmaskinene, systemene og protokollføre informasjon om testene. Om noen skulle prøve å etterligne det jeg har gjort bare med bakgrunn i opplæringen ville de ikke forstått det fordi de må også kunne min protokollføring på samme måte» Informant 3

«Alt jeg har lært, har jeg lært gjennom skole eller min forrige jobb. Der har de fortalt hvordan jeg skal beskrive min data på en god måte slik at andre forstår hva jeg har gjort og hva de betyr. Jeg har ikke fått noe opplæring i det her» Informant 4

Dataene viser hvordan mye av opplæringen er gjennomført på tidligere arbeidsplasser eller gjennom videreføring av andre ansatte. Videre er det tydelig hvordan opplæringen ikke er strømlinjeformet. De forskjellige ansatte har ansvaret for forskjellige testmaskiner eller testsystemer basert på stilling, og blir derfor opplært av andre med senior stillinger eller av noen som har brukt testsystemene før. Dette fører til en opplæring som ikke er strømlinjeformet og de ansatte kan jobbe med på veldig forskjellige måter. Videre kommer det frem i resultatene hvordan enkelte ønsker å digitalisere prosessene enda mer.

5.2.3 Databehandlingsprosessen

Resultatene viser hvordan databehandlingsprosessen er komplekst og det er mye som påvirker hva som kan gjøres med dataene, hvordan de lagres, hvordan dataen kontrolleres og hvem som har tilgang til dem. Det kommer også tydelig frem hvor viktig metadata er og hvordan det jobbes for å beskrive de forskjellige dataene.

Etter forskerne har hentet den nødvendige fra testmaskinen over til sin egen datamaskin, begynner jobben med kvalitetsjekking og plotting. Det kommer frem i resultatene hvordan all data har en verdi, selv om noen av dataene har større verdi enn andre. I dette case er de viktigste dataene de som skal plottes i en graf hvor dataene av mindre verdi er data som beskriver parameterne i testen som er gjennomført.

Når informantene blir spurt om hvordan de behandlet dataen kommer det frem hvordan de bruker Excel eller selvskrevne script i Python. De forteller også hvordan de selv har skrevet scriptene og de ikke har opplæring i dette. Noen velger å bruke scriptene til komplekse datasett, fordi det er enklere å behandle store data på denne måten istedenfor å bruke Excel. De velger denne tilnærmingen selv om de har utfordringer knyttet til bruk av Python og må gjennomføre workarounds for å få det til å gå opp. Det vises også hvordan Python har et større potensiale til å plotte data i større datasett enn Excel, selv om Excel er generelt enklere å jobbe i og derfor brukes i enklere datasett.

«Alle kan jobbe med Excel, men ikke all data kan brukes i Excel. Derfor bruker jeg Python for å plotte» Informant 4.

«All data har verdi, der noen er mer verdifulle enn andre. Jeg bruker outputen til å plotte en graf, og om jeg ser det er noe galt med grafen eller noen av verdiene sjekker jeg opp andre deler av datasettet for å se om det er en feil i testen. For å plotte grafen bruker jeg enten Excel eller et selvskrevet Python script. Som oftest bruker jeg Excel til mindre datasett, men til større datasett må jeg bruke Python.» Informant 1.

«Som oftest er jeg ute etter å lage en graf av dataene, litt mer sjeldent er det beregninger. Jeg har ingen konkret metode for å evaluere eller kvalitetssjekke. Det går litt automatisk på erfaring om hvordan jeg vet den skal se ut. Om jeg er usikker spør jeg noen som kan mer. Plotter dataene i Python eller Excel. Som sagt oftest grafer. Jeg bruker Python til større datasett. Jeg har erfaring gjennom MATLAB og kan derfor Python.» Informant 2

«Plotter mye i Python selv om jeg ikke har opplæring i dette og har laget et script som funker litt. Det er ikke optimalt, men det funker bedre enn Excel. Datasettene er så store at Excel har en tendens til å crashe. Jeg har en formening om hvordan resultatene skal se ut, og om de ikke er slik de skal er det noe feil. Om det er feil, må jeg finne ut av hvorfor og sjekke.» Informant 3

«Dataene jeg jobber med omhandler å beskrive andre data, og det er viktig for å forstå hva dataene egentlig betyr. Det bygges en database over slike beskrivelser og jeg er med å utvikle deler av denne.» Informant 6

I tillegg til hvilken metode informantene foretrekker å plotte sine data, kommer det frem hvordan de gjennomfører sine kvalitetssjekker av rådataene. Forskerne i denne institusjonen gjennomfører sine kvalitetssjekker basert på erfaring. De har en formening om hvordan de plottede grafene vil se ut, og om de avviker fra antagelsen, sjekker de om det en feil med testmaskinen, dataene som har kommet ut eller diskuterer med noen som har mer erfaring.

Videre kommer det frem i resultatene hvordan dataen deles og hvem som eier den produserte dataen. Ofte er dataene eid av den som produserer den og blir delt som grafer i presentasjoner eller gjennom en prosjektside i Microsoft Teams. Andre ganger kan virksomheten eller institusjonen som har betalt for prosjektet et ønske om å få tilgang til all rådataen fordi de ønsker å bearbeide denne selv.

Resultatene av data viser hvordan deling av data oftest gjennomføres gjennom avtalte løsninger som SharePoint eller gjennom GitHub. SharePoint brukes mest, og bruksområdet er lagring av rådata og plottede data. GitHub brukes til å dele data gjennom kode og i prosjekter hvor utvikling er en del av gjennomføringen.

«Ofte skal dataen jeg plotter være en del av et større prosjekt eller i en rapport. Derfor legger jeg grafen inn i en presentasjon vi samarbeider om og forklarer hva jeg har funnet, hva jeg har gjort og hvordan jeg har funnet den. Protokollen er med å beskrive hva jeg har gjort. Filnavnet korresponderer med forsøket og dato forsøket er gjennomført på. Mappestrukturen min er veldig strukturert, noe SharePoint ikke er. Dette gjør det vanskelig å finne ting i SharePoint.» Informant 1.

«Alt skal skje på SharePoint, men praksis og teori går ikke overens. Føler vaner hos folk er skylden her. I teorien skal alt være tilgjengelig og på et sted, og rådata lagres i sky. Dette er rotete i praksis. Jeg prøver å gjøre mine deler forståelig, men det er ingen praksis og alle gjør dette forskjellig. Det er begrensninger i hvordan SharePoint kan brukes og jeg skulle ønske de hadde en mer kompetent mappestruktur. Det blir veldig både og med deskriptive mapper.» Informant 2.

«Bruker Git for deling av kode gjennom grener, OneDrive for min egen del og Microsoft Teams for å dele dokumenter. Dataene er enten private, jeg kan be om tilgang eller så er de offentlige så jeg kan bruke de. Utfordringen med å dele data og kode er at brukerne må forstå hvorfor dataen kommer fordelt. Det tar tid å finne ut av hvordan og hvorfor dataen kommer slik, og det fører til frustrasjon og merarbeid. Derfor prioriterer jeg på å bruke mer brukte systemer som folk har kjennskap til» Informant 6

Det er stor enighet blant informantene om hvor stort behovet samhandlingssystemene er og viktigheten av bruken for å dele data og dokumenter, men gjennomføringen er ikke tilfredsstillende. Noen nevner hvordan det er vanskelig å finne frem i andres mapper, andre mener det fungerer bra i startfasen, men blir utfordrende og rotete etter hvert som det havner mer filer og mapper i de forskjellige kanalene. Det blir nevnt av et par informanter om hvordan de mener det er vaner som er skyldige og hvordan retningslinjer eller policyer for hvordan mapper og filer skal navngis vil være med å gi et bedre system. Andre mener det ikke vil ha noen innvirkning fordi kanalene i utgangspunktet er bygd ut av en standard med retningslinjer de skal følge, men retningslinjene ikke blir fulgt fordi de er mer til bry enn hjelp.

For å oppsummere databehandlingsprosessen viser resultatene hvordan forskerne i virksomheten bruker script for å plotte store datasett og Excel til å jobbe med mindre datasett. De har et stort ønske om å beskrive dataen på best mulig måte slik at andre skal finne de igjen i samhandlingsløsningene og forstå hva som er gjort, hva dataene betyr og hvordan de har jobbet for å få resultatene. Videre viser resultatene hvordan tanken bak samhandlingssystemene er bra i teorien, men utførelsen i praksis ikke er like god i flere plan.

6 Diskusjon

I dette kapittelet diskuteres funnene opp mot teorien for å kunne besvare forskningsspørsmålene:

F1: Hvordan kan forskere gjøre tilgang til forskningsdata enklere?
F2: Hvordan kan forskere gjøre forskningsdata mer forståelig?

Det er valgt å dele diskusjonen inn i to hoveddeler. I den første delen sammenlignes de to casene fra den multiple casestudien. Denne diskusjonsdelen vil beskrive likhetene og ulikhetene mellom de to casene, samt hvilke muligheter og utfordringer de har hver for seg og i felleskap.

I del to av diskusjonen vil utfordringene og mulighetene diskuteres opp mot den presenterte teoribakgrunnen for å finne forslag til hvor caseorganisasjonene burde rette fokus for å løse og minimere sine utfordringer eller gjøre tiltak for å kunne utnytte mulighetene. En del av denne diskusjonen vil innebære hvordan tidligere forskning beskriver databehandling som en strømlinjet prosess. Det vil diskuteres om forskningen presenterer databehandling på en riktig måte basert på hva som er presentert i resultatene.

6.1 Sammenligning av casene

I dette delkapittelet sammenlignes mulighetene og utfordringene i den multiple casestudien. Mulighetene og utfordringene diskuteres for å senere bli trukket opp mot det teoretiske grunnlaget for å kunne besvare forskningsspørsmålene.

6.1.1 Infrastruktur

De to forskjellige forskningsbyggene hvor forskerne med kontorarbeid eller labtester har forskjellige utfordringer og muligheter. I caset fra økologisk forskning kom det ikke frem noen utfordringer knyttet til fysisk infrastruktur. De ansatte virket ikke påvirket i stor grad av de ekstra sikkerhetstiltakene som er gjort og fordeling av testapparater så ut til å fungere i stor grad. Samtidig var testutstyret simpelt nok til å bruke caretakere i felt istedenfor forskere. Dette førte til hvordan forskeren kunne ha fullt fokus på å gjennomføre tester og arbeide med dataene istedenfor å måtte samle de inn i felt gjennom feltprøver. På en annen side, var den største utfordringen et byråkratisk problem fordi Miljødirektoratet ønsker å utvikle en unødvendig løsning fordi forskningsorganisasjonen allerede har et selvutviklet system som fungerer helt likt og er mulig å bruke til det samme formålet.

Fordelen med denne infrastrukturelle løsningen er å gi forskerne muligheten til å jobbe mer med data og testresultater. Det minimerer forskernes arbeidsoppgaver ved å bruke andre til å gjøre selve datainnsamlingen. Ulempen er å sikre riktig opplæring, ekstra økonomiske kostnader med å måtte lønne en caretaker og merarbeid med tanke på databehandlingen. I resultatene blir det nevnt hvordan en caretaker i noen felt må logge hvis hen går ut for å hente, endre eller sette opp et innsamlingsverktøy fordi det gjør utslag på testen. Selv om det hadde blitt gjort et utslag om forskeren var ute selv ville hen kunne logget når og visst når det loggete tidspunktet var gjennomført. Om systemet for logging ikke er godt nok er det mulig forskeren må inn i systemet og finne de eksakte

tidspunktene caretakeren var ute for å flagge deler av dataen som ugyldig på grunn av menneskelig påvirkning.

I motsetning til caseorganisasjonen i økologi eier ikke caseorganisasjonen i industriforskningen den mest brukt laben de brukte til undersøkelser. Av den grunn hadde de ikke helt styring på adgangskontroll og oppetid av internett. Hvor dette ikke var noe problem innen økologicaset er det en stor utfordring for industricaset. Tester som trenger flere hundre timer for å gjennomføres kunne bli avbrutt fordi eieren av bygget hvor laben ligger oppdaterte internettet eller slo det av for å gjennomføre endringer. En løsning vil være å gjennomføre hele testen ut tilkobling til nettverk, men det fører til en ny utfordring. Hvordan skal forskerne få tilgang til data uten internettilgang? Resultatene viser hvordan forskerne i industricaset bruker en ikke-kommersiell programvare for fjerntilgang. Denne løsningen vil ikke være mulig å bruke uten internett og fordi det ikke eksisterer noen policy for hvordan overføre data uten internett uten bruk av fysisk labbok hvor forskerne manuelt kan skrive ned resultatene for hånd blir en det skapt en loop av utfordringer uten noen klar løsning.

Følgelig, burde caseorganisasjonen se på hvordan disse utfordringene skal løses når den nye laben åpner for fullt. Åpningen av ny lab gir mange muligheter for å løse disse utfordringene om de riktige beslutningene blir tatt og de riktige personene får gi sin mening. Siden målet med den nye laben er å kunne samle alle data i en fil og gjøre dataen mer uniforme, vil det være nødvendig å gjøre gode valg for å sikre dette. Mengden data som blir samlet inn vil også øke drastisk. Derfor er det store muligheter for å kunne bli en pioner som kan drive forskningen fremover.

På en annen side vil dårlige valg kunne føre til ekstreme utfordringer. Det vil bli viktig å ta valg som løser de eksisterende problemene og samtidig må valgene være godt forankret i organisasjonen. Fordi organisasjonen eier den nye laben, vil ikke nødvendigvis internettilgangen være like stor utfordring, men det burde utarbeides tiltak i tilfelle det skjer. Samtidig burde de finne en løsning for hvordan de skal hente inn data fra en testmaskin som er av internett både i den eksisterende laben og den nye laben.

Ved å sammenligne de to casene er det tydelig hvordan det eksisterer utfordringer knyttet til infrastrukturer som kan påvirke kvaliteten i dataene som genereres. Selv om utfordringene er forskjellige vil infrastrukturelle utfordringer som å bytte et innsamlingsverktøy eller en internettoppdatering føre til endringer i data og merarbeid for forskerne. Dette er faktorer som påvirker innsamlingsprosessen i forskningen og forskerne må kontinuerlig tilbake til start for å gjennomføre arbeidsoppgaver igjen for å sikre høy kvalitet.

6.1.2 Opplæring

Da det kom til opplæring, var det veldig mye likt i resultatene. I begge casene er det tydelig hvordan begge organisasjonen har en lav prioritet om å lære bort like praksiser i organisasjonen. Selv om det er i gitt introduksjoner til systemer og testmaskiner, eksisterer det ingen overordnet policy i noen av organisasjonene på hvordan arbeidsoppgaver skal gjøres, prosedyrer skal beskrives eller data skal beskrives. Ved å se på utsagnet til informant 3 i 5.2.2, er det tydelig hvordan hen har fått opplæring av sin mentor, men ingen vil forstå hva som er gjort hvis de ikke har samme opplæring og protokollføring, noe som kan skape utfordringer.

Det kan være forskjellige årsaker til den manglende opplæringen i organisasjonene. Det kan ligge økonomiske begrunnelser til grunn som ikke gir muligheter til å sende

forskerne på opplæring, et ønske om å la forskerne føle mestring ved å lære seg selv å gjennomføre arbeidsoppgaver på sin egen måte eller uvitenhet rundt mulighetene opplæring vil kunne gi innad i organisasjonen. Uavhengig av hvilken begrunnelse som ligger til grunn skaper den manglende opplæringen utfordringer og et tap av muligheter.

Utfordringene som kan oppstå av den manglende opplæringen vil variere, men knyttet til forskningsspørsmålene vil lesbarhet og gjennomsiktighet i databehandlingsprosessen være en utfordring. Fordi opplæring er gjennomført av forskjellige enkeltpersoner i begge organisasjoner vil arbeidsmetodikk variere og det vil ikke være noen strømlinjet form for arbeidsvaner eller arbeidspreferanser. Selv om det kan være en fordel å la forskere og dataforskere jobbe på sin egen måte for å beskrive data og løse utfordringer vil det bli store forskjeller i hvordan oppgaver og tester gjennomføres. Om forskere jobber forskjellig, vil det kunne skape utfordringer med å forstå hverandres data om metadatabeskrivelsene ikke er gode nok.

En annen utfordring knyttet til opplæring er hvordan deltakerne er på få kurs. Ved å delta på kurs som en form for videreutdannelse eller etteropplæring, vil det skapes store muligheter for organisasjonen. I tillegg til påfyll av informasjon og opplæring i nyeste arbeidsmetodikker, kunne skape nettverk med andre som jobber på samme måten. Dermed kan kurs bli en nettverksbygger på linje med konferanser og ikke bare påfyll av informasjon.

Det er tydelig de to casene deler utfordringer og muligheter innen opplæring. Mangelen på opplæringen fører til forskjellige arbeidsmetodikker hos de ulike forskerne og dataforskerne, problemer knyttet til å forstå dataen hos de andre og et arbeidsmønster som ikke er strømlinjeformet og innebygd innad i organisasjonen. De nevnte utfordringene vil påvirke kvaliteten til data og hvordan forskere foretrekker å dele data med andre.

6.1.3 Databehandling

Databehandlingsprosessen i de to casene varierer fra rolle til rolle, noe som skaper likheter og ulikheter. For samarbeid i begge casene er bruken av Microsoft Teams og SharePoint veldig utbredt selv om holdningene til bruk varierer. I begge caser kommer det frem hvordan mye eller all kommunikasjon gjennomføres på Teams. Utfordringen ved bruk av Teams vises gjennom det menneskelige perspektivet. Det kom frem i resultatene fra begge caser hvordan forskjellige menneskelige vaner skaper rotete strukturer i praksis. I teorien føler informantene det er godt å ha alt på ett sted, men i praksis viser det seg hvor rotete kanalene er og ingen tar eierskap til dem. Denne utfordringen før til problemer med å navigere seg i løsningen og finne data eller filer som forskerne samarbeider om.

Denne utfordringen viser seg igjen frem i caset fra industriell forskning hvor det brukes GitHub til å dele data og kode. Fordi brukerne ikke nødvendigvis forstår hvorfor dataen kommer fordelt på den måten de gjør så vil det ta tid å navigere seg frem og forstå hvorfor dataen kommer på den måten. Dette fører til frustrasjon og merarbeid.

Selv om det er utfordringer knyttet til den infrastrukturelle mappestrukturen i Teamskanalene, er informantene enige om behovet for samhandlingssystemer for å kunne dele data og dokumenter. Det viste seg at den beste bruken av Teams for samhandling er deling av dokumenter, brainstorming og kommunikasjon i en mindre skala hvor noen føler eierskap til det som skjer istedenfor store felles kanaler hvor ingen har oversikt over hva som skjer.

På en annen side er det ulikheter når det kommer til hvordan casene kvalitetssjekker data og hvordan de arbeider med databehandling. Det er naturlige forskjeller her fordi informantene i de to casene har forskjellige roller. I caset med industriell forskning er det informanter som arbeider med data fra innsamlingstidspunktet til de er klare til å legges inn i en presentasjon eller deles med arbeidsgiver. Dette er annerledes hos økologi. Der er arbeider hovedsakelig dataforskere med kvalitetssjekker av ferdig bearbeidet data som sendes inn til databasen. Dermed blir arbeidet om å kvalitetssjekke data forskjellig.

I casen til økologisk forskning kommer ferdig bearbeidet data inn til databehandlerne eller forskerne får data inn gjennom felttestene. Testmaskinene utarbeider modeller og datasett basert på felttestene og forskeren gjennomgår dataene og sjekker validiteten. Et eksempel på hvordan det gjøres er å sjekke om en caretaker har vært ute og påvirket testresultatet før det eventuelt flagges som en hendelse. Når forskeren er ferdig med denne behandlingen blir datasettet overført til en database hvor de testes opp mot grenseverdier. Dataene som gir utslag mot grenseverdiene, blir flagget og datamanagerne sjekker om forskerne har beskrevet hvorfor eller ikke. Hvis det er en manglende beskrivelse tar en datamanager kontakt med forskeren som eier dataen gjennom problemløsersystemet for å rette opp i dette. Konsekvensen av denne tilnærmingen er flere. For det første sørger fremgangsmåten for en sikkerhet i kvaliteten på dataen og hvordan data som ligger i databasen er av god kvalitet. Samtidig setter fremgangsmåten krav til metadatabeskrivelser som beskriver dataen på en god måte slik at datamanageren forstår hva og hvorfor data blir flagget opp mot grenseverdiene.

På en annen side fører tilnærmingen til mye dobbeltarbeid og muligheter for menneskelige feil. En av informantene forteller om hvordan hen ønsker å redusere menneskelige feil og redusere mulig dobbeltarbeid ved å automatisere prosessene rundt kvalitetssjekk. Dette er en mulighet i organisasjonen da de sitter på ressursene og kunnskapen som trengs.

Hos caset i den industrielle forskningen er det en annen tilnærming til kvalitetssjekk. I caset er all data verdifull selv om noe er mer verdifullt enn andre. De mest verdifulle dataene er resultatene av en test og de minst verdifulle er dataene som beskriver testutstyret. Etter at data er plottet gjennom Python eller Excel avhengig av datasettets størrelse gjennomføres det en kvalitetssjekk. Kvalitetssjekken gjennomføres på mageølelse og en antagelse om hvordan den ferdige grafen kommer til å se ut. Altså hvordan de viktige dataene korresponderer med hverandre. Om forskeren får en graf som ser uriktig ut blir denne diskutert med andre som har erfaring innen området eller så gjennomføres en kontrollsjekk på de mindre viktige dataene som beskriver testmaskinen i tilfelle en feil har oppstått.

Utfordringen med denne tilnærmingen er som i forrige case menneskelige feil. I dette caset er det identifisert to hovedområder hvor menneskelige feil kan oppstå i prosessen. Den første er i store datasett hvor forskeren bruker selvlærte Pythonscript for å plotte en graf, mens den andre utfordringen er kvalitetssjekken på den plottede dataen. Fordi bare dataene av høy verdi sjekkes kan det være data av lav verdi som kan ha hatt en påvirkning som gir et lite avvik. Det er også mulig en kombinasjon av mageølelse og et selvlært Pythonscript kan påvirke kvaliteten.

For begge casene er det viktig å beskrive data på en god måte. Selv om det er nevnte utfordringer til arbeidsflyt og forskjellige metoder er ønsket å om å beskrive data på en universell måte for å skape forståelse og muligheter for gjenbruk stort. Derfor legger forskerne store ressurser i å beskrive dataene og i casen fra industriforskning drives det

også arbeid for å bygge en database over slike beskrivelser. Dermed er det tydelig hvor viktig beskrivelser for å danne forståelse er viktig i begge casene.

Når det kommer til deling av data, er det noen forskjeller og noen likheter. For begge casene eksisterer det regler for hvem som eier dataene og hvem som kan bruke de vederlagsfritt noe som føles trygt. I begge casene er det hovedsakelig oppdragsgiver som bestemmer om dataen skal publiseres eller ikke. Fra den økologiske casen kom det frem hvordan deres type forskning er tidlig ute med å dele data for å drive forskningen fremover og kunne diskutere med andre. Mye av oppdragene som ble gjennomført av dem er med offentlig oppdragsgiver. Dermed er holdningen å publisere offentlige finansierte data slik at den er tilgjengelig for alle. Med bakgrunn i dette blir holdningen for organisasjonen hvordan de ikke eier dataen, de distribuerer den og gir tilgang gjennom sin database.

På den andre siden, er det forskjell hvis oppdragsgiveren er ikke er offentlig noe som gjenspeiler seg i industricaset. Om det er oppdrag hvor industri er arbeidsgiver blir det utarbeidet avtaler om hvem som eier dataene og hvem som skal ha tilgang. Dette gjelder for begge casene. For industricaset er det enda en faktor som spiller inn. Hvilke data ønsker de egentlig å dele? Industriforskningen er drevet av muligheter for patenter og konkurranse mot andre. Ved å dele data kan utfordrere komme foran i forskningen og komme først frem til en løsning eller et svar som kan skaffe et patent.

6.1.4 Oppsummering av sammenligningen

For å oppsummere deler de to casene noen utfordringer og muligheter. Samtidig har de noen utfordringer og muligheter den andre ikke har. Innen infrastruktur har industricaset store muligheter til forbedringer og kan gjøre tiltak på den nye laben som vil kunne endre arbeidsmetodene og arbeidsstrømmen til de ansatte. Samtidig som de har muligheten til å utbedre og løse de infrastrukturelle utfordringene knyttet til datainnsamlingen.

Videre har begge casene utfordringer knyttet til opplæring. Utfordringene skaper forskjellig arbeidsstrøm, lite arbeidsflyt og metoder for kvalitetssjekk og menneskelige vaner som påvirker bruken av samhandlingsverktøy. Her eksisterer det muligheter til å iverksette tiltak som vil gjøre arbeidsmetodene mer strømlinjeformet, gi opplæring i forskjellige metodikker å jobbe på samt sette retningslinjer for hvordan beskrive data.

Til slutt har begge caser noen utfordringer og muligheter til å dele data. Selv om det eksisterer regler og det ene caset har utfordringer med å bestemme hvilke data som skal deles, har begge casene samhandlingssystemer og samhandlingsmetoder som gjør det mulig å presentere og dele data med andre.

6.2 Utfordringer og muligheter satt opp mot teoretisk bakgrunn

I dette delkapittelet blir utfordringene og mulighetene diskutert opp mot det teoretiske bakgrunnen presentert i kapittel 2. Delkapittelet beskriver hvordan muligheter opplæring gir organisasjonen fra casene og hvilke utfordringer dette kan være med å løse. Videre diskuteres hvilken tilnærming forskning har til databehandling og hvilke muligheter og utfordringer en ny innfallsvinkel vil gi. Til slutt avsluttes delkapittelet med en oppsummering av de to diskusjonene.

6.2.1 Muligheter med opplæring

Som nevnt tidligere, er det utfordringer og muligheter i begge casene som påvirker hvordan forskningsspørsmålene kan besvares. En løsning som kan være med å løse mange av utfordringene er å øke ressursbruken og fremme viktigheten av opplæring i organisasjonene. Selv om det gjennomføres opplæring i bruk av systemer, innsamlingsverktøy og testverktøy er det et stort potensiale til å gjennomføre opplæring innen datalagring og prosessering, kvalitetssjekk av data og deling av data.

Opplæring innen datalagring og prosessering kan gi store muligheter for organisasjonene, spesielt innen industriell forskning caset hvor de selv er ansvarlige for hele prosessen. For de store mengdene data føler forskeren Excel er et dårlig verktøy, noe som gjenspeiler seg i utsagnet fra Borer et al. (2009) om hvordan regnearkprogram vil byttes ut ved andre. Basert på funnen i studien er det programmer som er kompatible med Python forskerne velger å bruke på grunn av de store mengdene data de behandler. Skript kan brukes på flere måter og resultatene kan en enkelt presenteres på en god måte (Vliet, 2019). Fordi et stort flertall av forskerne ikke har utdanning som tilrettelegger for bruk av Python har de lært seg dette på egenhånd. Ved å gjennomføre opplæring for forskere som behandler data og prosesserer data, vil organisasjonene sikre bedre skript for presentering av resultater og kunne øke kvaliteten fordi det gir opplæring i et skript som kan brukes til å plotte data.

Videre kan opplæring gi muligheter innen kvalitetssjekker og det kan være med å bedre metadatabeskrivelser og sikre lesbarhet i dataene. I begge casene er gjennomføres det kvalitetssjekker. Selv om de har forskjellige retningslinjer for hvordan, kan prosessen forbedres. Der kvalitetssjekken var en avgjørelse basert på mageføelse eller et utslag mot en grenseverdi sjekkes i realiteten et utvalg elementer fra Cai & Zhu sin to lags modell for Big Data kvalitetsstandard (Cai & Zhu, 2015). Elementene som inngår i kvalitetssjekken er nøyaktighet, kompletthet, om de er i samsvar med forskerens søkningstema og om dataene er innenfor det akseptable omfanget av verdier. Ved å lære opp forskerne til å forstå viktigheten av resten av elementene og dimensjonene presentert i Figur 2 (Cai & Zhu, 2015), vil det kunne endre retningslinjer og gjennomføring kvalitetssjekkene. Samtidig vil det føre til data som er enklere å forstå og det vil være klassifiseringer og beskrivelser av dataene som tilfredsstiller flere forskere sitt behov under bruk.

Andre utfordringer opplæring kan være med å løse, er bedre retningslinjer for hvordan bruke Teams og SharePoint i organisasjonene. I caset til økologisk forskning brukes Teams og SharePoint hovedsakelig som kommunikasjonskanaler og noe samskriving mens det i industriell forskning i tillegg brukes som et depot hvor data lagres. I begge caser beskrives løsningene som kaotiske på grunn av hvordan menneskene bruker dem. Ved å legge retningslinjer for bruk og følge dem, vil forskerne lettere kunne få gjennomført privat deling av data i sky og det vil bli bedre kommunikasjonskanaler. Allagnat et al., (2019) sin forskning beskriver hvordan deling av data i skyløsninger gjennomføres når det er privat data som deles. I samarbeidsprosjekter hvor data deles privat mellom forfattere av en rapport eller data ikke ønskes å deles offentlig kan dermed Teams/SharePoint fungere som et datadepot om det legges retningslinjer forskerne følger.

Samtidig kan opplæring være med å forskere en bedre forståelse av viktigheten av metadata, hvilke standarder som eksisterer og hvordan de kan benytte dem for å sikre lesbarhet i sin data. Som nevnt finnes det flere metadataverktøy hvor over 40 av dem er

brukt eller nevnt i vitenskapelige manuskript fra 1997 til 2018 (Alves et al., 2018). Borer et al. (2009) og Alves et al. (2018) nevner begge EML som en utbredt standard som forskere kan benytte for metadatabeskrivelser. Ved å presentere og gi forskerne innsikt i hvordan benytte slike vil dataene de samler inn bli bedre beskrevet og være mer lesbar om de deles. For den industrielle forskningen, hvor det jobbes med ontologier, vil det også være viktig å ha en forståelse for hvilken plattform forskerne ønsker å bruke. Selv om Alves et al. (2018) sin forskning er gjort innenfor økologisk forskning, viser denne studien hvordan det er likheter mellom casene og derfor vil det være like utfordringer når det kommer til integrering av dette metadataverktøyet og hvordan bruke det i et forskningsprosjekt.

Til slutt kan opplæring være med å gi forskerne et innblikk i hvordan de byråkratiske reglene er til hjelp, samt få en bedre forståelse for hvorfor og hvordan data kan deles. Som beskrevet tidligere, differensierer casene noe når det kommer til ønsket om å dele data. Selv om begge casene er enige om deling av data vil drive feltet fremover, er det konkurranse om patenter hos industriell forskning og majoriteten av den økologiske forskningen er offentlig og skal derfor som hovedregel deles. Ønsket om å dele data sammenfaller med forskningen til Allagnat, hvor forskere i Japan ønsker å dele dataene fordi det kan være viktig for andre og det motiverer å kunne ha en økende fremgang i feltet (Allagnat et al., 2019). Samtidig gjenspeiler utfordringene Karasti og Allagnat beskriver. For forskerne er det en sosial kompleks prosess der de må balansere press, interesser og tanker om misbruk av data (Allagnat et al., 2019; Karasti et al., 2006)

Selv om forskerne i casene og fra forskningen til Karasti et al. (2006) velger å dele data fordi det er flere fordeler enn ulemper er det forskjeller til hvordan dataen deles. Databasen caset til industriell forskning benytter blir beskrevet som en god løsning av Hrynaszkiewicz et al. (2001) som en god måte å lagre dataene på. Videre mener de opplæring kan gi ferdigheter og forståelse om viktigheten av enkle praksiser som å lagre data i depoter eller databaser (Hrynaszkiewicz et al., 2021). På en annen side er ikke denne løsningen integrert med den tradisjonelle publiseringen. Dette kan være et steg videre for å komme nærmere en optimal løsning for datadeling da 51% av forskere fra PLOS allerede deler data som et supplement til journalartikler (Allagnat et al., 2019; Hrynaszkiewicz et al., 2021). Ved å gi forskere denne opplæringen vil de forstå viktigheten av å dele data som er forståelige med hverandre for å drive forskningen fremover og kunne gjør nye fremskritt.

På en annen side er det utfordringer knyttet til opplæring. Utenom de faktiske økonomiske kostnadene knyttet til å gjennomføre opplæring enten internt eller eksternt, er det andre faktorer som spiller inn. Selv om det er et ønske fra flere om å ha en gjennomarbeid og strukturert arbeidsflyt internt i organisasjonen, vil opplæring gjøre denne arbeidsflyten påtvunget og for noen forskere vil dette føles ut som en påtvunget endring som innebærer en endring i atferdsmønster på jobb, hvor de for eksempel må bruke Python som skriptspråk for større datasett da opplæring er gitt i det programmeringsspråket.

En annen kostnad som kan påvirke organisasjonen er å ta forskerne ut av sin stilling for å gjennomføre opplæring. Forskerne vil dermed bruke arbeidstiden sin på opplæring istedenfor sin forskning. Dette vil skape kostnader i form av tapt arbeidstid eller andre som må fylle inn for den som er borte. Alternativet til opplæring er å bruke tid på gruppemøter for å prate om utfordringene og viktigheten av tiltakene opplæring kan gi.

6.2.2 Et nytt syn på databehandling

Fra teorien blir databehandlingsprosessen presentert som en strømlinjeformet prosess hvor data samles inn, prosesseres, lagres og deles. Fra resultatene og observasjonene kom det frem hvordan denne prosessen ikke er like strømlinjeformet som det blir presentert i forskningsartikler. Tidligere forskning beskriver mulighetene for innhenting av data og hvordan enorme mengder som Big Data skaper utfordringer knyttet til hvilken grad denne dataen er av kvalitet og hvilke utfordringer som er knyttet til å sikre høy kvalitet (De Mauro et al., 2015; Diepenbroek et al., 2014; misqresearchcurations, 2022). I teorien blir kvalitetssjekk sett på som en fase som kommer etter prosessering og lagring, men det viser seg kvalitet spiller inn allerede før dataene er samlet inn. Fordi utviklingen av datainnsamlingsmetoder har gitt enorme muligheter de siste årene må forskere allerede før dataene er samlet inn gjøre tiltak og forsikre seg om kvaliteten på dataene som kommer inn. Denne utfordringen gjelder også i casene. I økologisk forskning er det allerede beskrevet problemer knyttet til kvalitet på datainnsamlingen når de skifter testutstyr som påvirker kvaliteten til målingene.

Videre er datalagring og prosessering en kontinuerlig prosess. Det er beskrevet hvordan forskningsdata innen økologi vanligvis lagres i regneark eller lokale databaser (Diepenbroek et al., 2014), noe som har gjenspeilet seg i funnene fra begge caser. Selv om det er beskrevet hvordan regnearkverktøy er på vei ut og nye systemer vil ta over (Borer et al., 2009), viser det seg det bare gjelder større datasett som Excel ikke klarer å håndtere. Dermed er det en kontinuerlig oppdatering av regnearkene og metadatabeskrivelser hvor forskere har muligheten til å finne igjen tidligere data og sammenligne dem med nyere data for å beskrive fenomener som har skjedd. Igjen vil det være viktig å sikre kvalitet i gammel data og ny data for å kunne ta gode avgjørelser.

Videre er det forskjellige tidspunkter for når data deles. Funnene viser hvordan det noen ganger deles rådata, noen ganger deles ferdig prosesserte i form av plottede grafer eller den kan legges ved til publikasjoner som rådata. Dette gjenspeiler seg i forskningen til Allagnat et al., (2019) og Hrynaszkiewicz et al., (2021). Denne formen for data viser hvordan det hopper fra å dele data før den er prosessert og lagret til etter den er prosessert og lagret.

Ved å se eksemplene over, kommer det frem hvordan databehandlingsprosessen ikke er like strømlinjeformet som presentert i tidligere forskning. Prosessen er mer kompleks, og de forskjellige stegene vil påvirke hverandre og til slutt kvaliteten i dataene. Ved å se på databehandling som en kontinuerlig prosess hvor alle fasene har en tilknytning til hverandre, vil forskningen kunne gjøre store steg fremover. Et eksempel er hvordan teamet kvalitet må få en ny innfallsvinkel. Som nevnt tidligere, er en av utfordringene knyttet til store menger data hvordan det er vanskelig å bedømme datakvalitet innen rimelig tid og hvordan data og aktualitet endres raskt (Cai & Zhu, 2015). Andre utfordringer som må utforskes med en innfallsvinkel der alle fasene har en sammenheng er som De Mauro et al., (2015) beskriver; hvilken datainnsamling er av kvalitet?

Videre er presentasjonskvalitet allerede et tema (Cai & Zhu, 2015). Skript gir muligheter for å presentere data (Borer et al., 2009; Vliet, 2019) ferdig prosessert, mens gode metadatabeskrivelser vil være med å gi kvalitet og mening i data egentlig betyr (Alves et al., 2018; Borer et al., 2009; Diepenbroek et al., 2014; Gjersdal & Nätt, 2023). Ved å se på databehandling fra dette ståstedet åpner det seg mange utfordringer som må løses, men også mange muligheter som vil være verdt å utforske.

6.2.3 Oppsummering

For å oppsummere kan har opplæring potensialet til å kunne løse utfordringene som blir presentert i casene og gi ut grunnlag for å kunne benytte seg av de forskjellige mulighetene som er presentert. Samtidig er det utfordringer knyttet til opplæring som vil påvirke hvordan de ansatte jobber og spørsmålet om hvor lønnsomt opplæring er i organisasjonen må utredes før en avgjørelse kan tas.

For å løse utfordringene er det viktigste hvordan de blir adressert. Uavhengig om det blir gjennomført opplæring eller ikke må forskerne bli bevist på hvilke utfordringer de står ovenfor i sin arbeidshverdag og hvilke verktøy de har for å kunne løse utfordringene og gjennomføre databehandlingsprosessen på en god måte for å kunne sikre høy kvalitet i data og muligheter for å dele denne med andre interessenter.

Samtidig er det tydelig hvor komplekst det er å gjennomføre databehandling innen forskning. Det er mange utfordringer knyttet til innsamling, prosessering, lagring, kvalitetssjekk og samspillet mellom de forskjellige fasene i prosessen. Forskere må forstå hvordan databehandling påvirker deres arbeid og hvordan de må jobbe med dataene for å skape gode resultater.

I oppgaven er det presentert to forskningsspørsmål. For å besvare F1, «Hvordan kan forskere gjøre tilgang til forskningsdata enklere?», burde forskere se på muligheten for å skape databaser eller datadepoter med adgangsstyring. Det er diskutert tidligere hvordan denne tilnærmingen er teoretisk forankret (Allagnat et al., 2019; Hrynaszkiewicz et al., 2021), og begge caser har en løsning hver for dette. Den økologiske forskningen bruker en database og den industrielle forskningscasen benytter seg av Teams med integrert SharePoint som et datadepot. Selv om begge bruker en løsning som er teoretisk forankret, har de begge forskjellige utfordringer. Databasen kan ha utfordringer med ugyldig data som kommer inn fra forskere, mens datadepotet til det industrielle caset har utfordringer knyttet til menneskelige bruk og vaner. Databehandling er som diskutert en kompleks prosess og derfor er det vanskelig å finne en perfekt løsning. Derfor burde caseorganisasjonene vurdere om de burde gjøre endringer til sine løsninger for å styrke databasene eller datadepotene.

Videre prøver oppgaven å besvare F2, «Hvordan kan forskere gjøre forskningsdata mer forståelig?» I caset til økologisk forskning benytter de seg av flagg for å beskrive data som er av forskjellige årsaker når grenseverdier. På den andre siden, er det vanskelig for andre å forstå hva dataene betyr i det industrielle caset om de ikke har fremgangsmetoden og beskrivelsene fra forskerne. I det teoretiske grunnlaget blir det presentert et mangfold av muligheter for metadatabeskrivelser gjennom ontologier eller standarder som EML (Alves et al., 2018; Borer et al., 2009). I begge casene burde de gjøre tiltak for å gjøre forskningsdata mer forståelig. Industricaset jobber med å bruke ontologier for å kunne beskrive dette. Det er en tilnærming som kan fungere. Det viktigste vil være å benytte seg av en standard for metadatabeskrivelser som fungerer i organisasjonen og er anerkjent nok til at den kan legges ved datadelingen. Dermed vil forskere som får tilgang til dataen kunne få en forståelse for hva den betyr fordi det ligger ved en beskrivelse i form av ontologier eller en standard som er forståelig.

7 Konklusjon

I dette kapittelet gjennomføres en konklusjon av oppgaven. Kapittelet starter med en gjennomgang funnene. Deretter beskrives oppgavens begrensninger og det blir gitt noen forslag til videre arbeid. Kapittelet avsluttes med hvordan prosjektet relaterer til FN sine bærekraftsmål.

Det er funnet flere muligheter og utfordringer knyttet til databehandling innen forskningsmiljøer i prosjektet. Gjennom casene er det funnet infrastrukturelle utfordringer som kan være med å påvirke datainnsamlingen. Det er også identifisert tiltak organisasjonene kan gjøre for å minimere eller løse utfordringene. Spesielt caset for industriell forskning har en kjempemulighet til å skape en plattform for datainnsamling med den nye laben. Hvis caseorganisasjonen utreder de eksisterende utfordringene ved den nåværende laben og legger ressurser i å et godt forarbeid, kan den nye laben bli ledende på sitt felt.

Begge caser har også utfordringer og muligheter knyttet til databehandling. Det er tydelig hvordan Excel som databehandlingsverktøy har sine begrensninger ved større datasett. Mye av databehandlingen gjennomføres enten i dette verktøyet eller blir plottet med Python selv om majoriteten av forskere har liten erfaring med programmeringsspråket fra før. Denne utfordringen fører til merarbeid og lite effektiv bruk av arbeidstid.

Videre har de to casene forskjellige holdninger og retningslinjer til hvordan kvalitetssjekke og beskrive data. I caset for økologisk forskning gjennomføres kvalitetssjekk på to måter. En automatisert prosess sjekker dataen forskerne sender inn opp mot grenseverdier, og verdier for nære eller over grensen blir flagget. Forskerne har også muligheter til å flagge dataene selv på forhånd. Deretter sjekkes de flaggete dataene manuelt og om metadatabeskrivelsene er for dårlige vil dataforskeren kontakte forskeren som sendte inn dataene for å løse problemet. Den industrielle forskningen, benytter seg av mageølelse og erfaring etter å ha plottet dataene for å besvare det samme spørsmålet.

Deling av data gjennomføres i begge caser hovedsakelig gjennom datadepoter eller databaser. Denne tilnærmingen til deling av data blir beskrevet av tidligere forskning som det beste måten å dele data på (Allagnat et al., 2019; Hrynaszkiewicz et al., 2021). Selv om begge har utfordringer knyttet til hvordan data deling fungerer i praksis benytter de seg av gode løsninger.

Samtidig kan funnene være med å bestride den strømlinjeformede beskrivelsen tidligere forskning har av databehandling (Bordelon, 2023; Borer et al., 2009; misqresearchcurations, 2022). Forskningen beskriver denne prosessen som en enkel strømlinjet prosess hvor alt skjer etter hverandre. Gjennom diskusjonen beskrives hvordan denne databehandling er mer komplekst enn hva tidligere forskning har kommet frem til. Delprosessene i databehandling har en større påvirkning på hverandre enn det som beskrives og måten forskere ser på databehandling, spesielt innen forskning, burde diskuteres.

For å konkludere og besvare forskningsspørsmålene er det mange muligheter og utfordringer knyttet til databehandling i forskningsmiljøene. Begge casene har en tilnærming til deling av data som er knyttet opp mot den beste løsningen fra tidligere forskning. For å besvare hvordan forskere kan gjøre forskningsdata mer tilgjengelig (F1), er det viktig for casene å forstå hvilke utfordringer de har og hvordan de kan utrede disse. Utfordringene er forskjellig, men hovedtanken med å ha en database eller et datadepot hvor de kan ha adgangsstyring ovenfor forskere viser seg å være den optimale løsningen.

Samtidig må forskerne på institusjonene bestemme seg for hvilken metadataløsning de ønsker å benytte seg av. For å gjøre forskningsdataene mer forståelig (F2), Må casene benytte seg av et anerkjent metadataverktøy eller standard som EML. Ved å benytte seg av en standard som er kompatibel med innsamling og databehandlingsverktøyet caseorganisasjonene benytter seg av, vil de kunne legge ved denne til datadelingen og dataene som blir delt blir dermed bedre beskrevet og forklart.

7.1 Begrensninger og videre arbeid

Oppgaven er begrenset til to forskjellige forskningsinstitusjoner som har forskere i forskjellige felt. Casevirksomheten beskrevet i 3.1 gir tilgang til forskning innen luftforskning og en database hvor det lagres atmosfærisk komposisjonsdata fra flere nasjonale og internasjonale rammeverk, samt andre forskningsprosjekter. Organisasjonen beskrevet i case 3.2 har gitt tilgang til sine laber innen hydrogen og batteri. Dermed er forskerne som har stilt som frivillige i datainnsamlingen begrenset til å jobbe innen de nevnte forskningsmiljøene og institusjonene.

Videre er det gjort begrensninger knyttet opp til hva databehandling innebærer. I oppgaven er databehandling definert som hvordan forskere samler data, lager datasett, sjekker kvalitet på data, evaluerer data og hvordan de deler sine funn med andre forskere. Dermed vil former for databehandling som ikke inngår i definisjonen ikke bli diskutert i oppgaven.

En annen begrensning er hvilke data som forskerne har hentet inn. I oppgaven begrenses datainnsamling hos forskerne til datainnsamling de selv har gjennomført eller de har betalt noen til å gjennomføre. Data som genereres som brukerdata fra apper og andre internettbesøk er dermed ikke en del av denne oppgaven fordi det ikke er naturlig for noen av informantene å bruke denne typen data i sin forskning.

Basert på begrensingene som er gitt ved oppgaven anbefales videre forskning å utforske andre forskningsmiljøer for å se på flere forskjeller og likheter som eksisterer på tvers av miljøene. Samtidig burde forskningen se på hvilke forskjeller som eksisterer i utlandet. Oppgaven er basert på to norske institusjoner og med bakgrunn i kulturforskjeller kan funnene variere i andre land.

Videre burde casene ta de adresserte utfordringene og mulighetene til rette. I begge caser er det utfordringer som kan løses og muligheter som burde utforskes for å fremme gid datadeling av kvalitet. Selv om oppgaven diskuterer opplæring som et verktøy for å løse dette, er det viktig for organisasjonene å finne sine løsninger på de adresserte funnene. Dermed burde videre arbeid for forskningsinstitusjonene å sette ned en prosjektgruppe som kan finne de beste løsningene.

Til slutt burde forskere vurdere å starte et prosjekt som ser på om databehandling er den perfekte strømlinjeformede prosessen de selv beskriver i sine rapporter. Det kommer

tydelig frem fra funnene i dette prosjektet hvordan databehandling er en større kompleks prosess en det som er beskrevet i forskningen. Derfor burde dette utredes for å få en bedre forståelse for hvordan databehandling fungerer i forskning og muligens også innen industri.

7.2 Bærekraft

Når det kommer til bærekraft, er oppgaven direkte relatert til bærekraftsmål nummer 9 hos FN. Målet er å: «bygge solid infrastruktur og fremme inkludere og bærekraftig industrialisering og innovasjon (FN-sambandet, 2023).» Utfordringene forskere har knyttet til datainnsamling, databehandling og datalagring vil trekkes opp mot dette feltet. Spesielt delmål 9.5 hvor det fokuseres på å styrke vitenskapelig forskning og betydelig øke bevilgningene til offentlig og privat forskning, og delmål 9.b hvor det skal støttes nasjonal utvikling av teknologi og forskning (FN-sambandet, 2023). Ved å løse de utfordringene og ta mulighetene som blir adressert i diskusjonen vil de to organisasjonene være med å hjelpe Norge nå bli bedre innen dette feltet. En av utfordringene Norge har er hvordan gjøre industrien mer bærekraftig og derfor må det investeres mer i forskning og teknologi (FN-sambandet, 2023). Begge casene har muligheten til å øke bærekraften i egen forskning, men spesielt caset for industriell forskning har en god mulighet til å hjelpe Norge ved å nå dette målet.

Referanser

Allagnat, L., Allin, K., Baynes, G., Hrynaszkiewicz, I., & Lucraft, M. (2019, mai 22). *Challenges and Opportunities for Data Sharing in Japan* [Online resource]. Figshare; figshare. https://doi.org/10.6084/m9.figshare.7999451.v1

Alves, C., Castro, J. A., & Ribeiro, C. (2018). *Research data management in the field of Ecology:an overview*. 8.

Baxter, P., & Jack, S. (2015). Qualitative Case Study Methodology: Study Design and Implementation for Novice Researchers. *The Qualitative Report*. https://doi.org/10.46743/2160-3715/2008.1573

Bordelon, D. (2023, februar 2). *Guides: Research Data Management @ Pitt: Understanding Research Data Management*. Research Datamanagement @ Pitt. https://pitt.libguides.com/managedata/understanding

Borer, E. T., Seabloom, E. W., Jones, M. B., & Schildhauer, M. (2009). Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America*, *90*(2), 205–214.

Cai, L., & Zhu, Y. (2015). *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era* (Nr. 0). *14*(0), Artikkel 0. https://doi.org/10.5334/dsj-2015-002

De Mauro, A., Greco, M., & Grimaldi, M. (2015). *What is big data? A consensual definition and a review of key research topics*. 97–104. https://doi.org/10.1063/1.4907823

Diepenbroek, M., Glockner, F. O., Grobe, P., Konig, Z. F., Guntsch, A., Garten, B., Huber, R., Konig-Ries, B., Jena, U., Kostadinov, I., Unversity, J., Nieschulze, J., Seeger, B., Tolksdorf, R., & Triebel, D. (2014). *Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)*. 11.

FN-sambandet. (2023, januar 31). *Industri, innovasjon og infrastruktur*. https://www.fn.no/om-fn/fns-baerekraftsmaal/industri-innovasjon-og-infrastruktur

Gjersdal, A., & Nätt, T. H. (2023). Metadata. I *Store norske leksikon*. https://snl.no/metadata

Helmholtz, D. K., & Fraunhofer, U. J. (2022). *DESCA Model Consortium Agreement—DESCA 2020 Model Consortium Agreement*. https://www.desca-agreement.eu/desca-model-consortium-agreement/

Hrynaszkiewicz, I., Harney, J., & Cadwallader, L. (2021). A Survey of Researchers' Needs and Priorities for Data Sharing. *Data Science Journal*, *20*(1), Artikkel 1. https://doi.org/10.5334/dsj-2021-031

Karasti, H., Baker, K. S., & Halkola, E. (2006). Enriching the Notion of Data Curation in E-Science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network. *Computer Supported Cooperative Work (CSCW)*, *15*(4), 321–358. https://doi.org/10.1007/s10606-006-9023-2

Krumsvik, Rune Johan. (2015). *Forskningsdesign og kvalitativ metode. Ei innføring* (2. utg.). Fagbokforlaget Vigmostad & Bjørke AS.

misqresearchcurations. (2022, februar 14). *Data Management*. MIS Quarterly. https://www.misqresearchcurations.org/blog/2022/2/11/data-management

Oates, B., Griffiths, M., & McLean, R. (2022). *Researching Information Systems and Computing* (2nd Edition). SAGE.

Sander, K. (2021, juli 21). *Organisasjonen som et åpent sosio-teknisk system*. eStudie.no. https://estudie.no/organisasjonen-apent-sosio-teknisk-system/

Tjora, A. (2021). *Kvalitative forskningsmetoder i praksis* (4. utg.). Gyldendal.

Vliet, M. (2019). *Guidelines for data analysis scripts*.

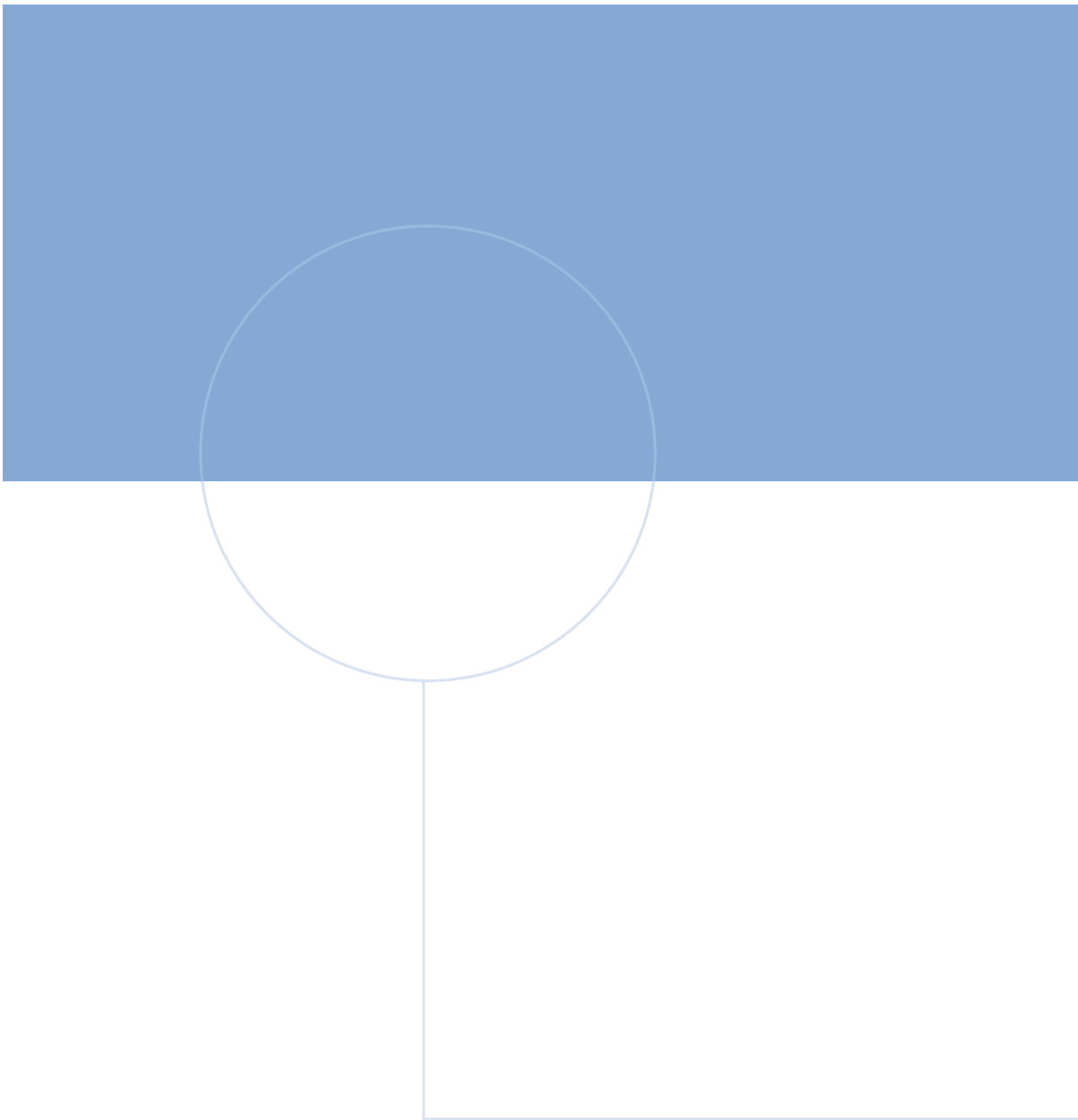
Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems 12*, *4*, 5–33.

Vedlegg

Vedlegg A, Intervjuguide.

Tema	Spørsmål
Intro/generelt	Hva er din rolle i prosjektet? Hvor lenge har du jobbet i dette forskningsmiljøet? Har du jobbet i andre forskningsmiljøer tidligere?
Opplæring i databehandling	Hvordan opplæring har du med databehandling fra studie? Hvordan opplæring har du med databehandling fra din arbeidsplass?
Samarbeid og deling av data	Hva tenker du på når jeg sier samarbeid? Mener du det er mest naturlig for deg å jobbe uavhengig av andre eller sammen for å løse en oppgave? Hvilke teknologier bruker du til samarbeid nå? Hvilken bruker du mest og hva bruker du den til? Endres teknologien for samarbeid ofte? Endres teknologien for innsamling av data ofte? Hvordan deler du dine data i dag? Hvordan deler du dine funn i dag? Har din deling av data og funn endret seg etter du startet å arbeide? Kan du forklare hvordan? Hvilke teknologier bruker du til å dele data i dag? Hvordan bruker du teknologiene? Hvordan opplever du teknologien hjelper deg i prosjektet? Hvilken nytteverdi opplever du teknologien har?
Data fra andre	Hvordan jobber du med data andre har hentet inn? Kan du fortelle litt om hvordan du jobber for å forsikre deg om at dataen er av kvalitet?
Byråkrati	Er du pålagt å gjøre data tilgjengelig for andre forskningsstasjoner? i) Hvis ja, hvem stiller disse kravene? Det finnes regler om hvem som eier data og hvem som kan bruke dem vederlagsfritt. Hva tenker du om slike regler? Er de

	betryggende, eller føler de legger mer restriksjoner enn de er til hjelp?
Mer om deling av data	Kan du forklare noen aspekter med samarbeid og deling av data du liker? Kan du forklare noen aspekter med samarbeid og deling av data du ikke liker? Har du noen tanker om hvordan samarbeidet og deling av data kunne blitt bedre?
Avslutning	Er det noe mer relevant du kan fortelle oss deling av data i arbeidet ditt? Er det noe mer relevant du kan fortelle om datakvalitet i arbeidet ditt? Er det noe mer relevant du kan fortelle om samarbeid i arbeidet ditt?



NTNU

Kunnskap for en bedre verden