HØGSKOLEN I OSLO
OG AKERSHUS

Kristine Aalrust Kristoffersen

Mapping med mening

Utfordringer ved mapping av indekseringsspråk

Masteroppgave 2015
Master i bibliotek- og informasjonsvitenskap
Høgskolen i Oslo og Akershus, Institutt for arkiv- bibliotek- og informasjonsfag

Sammendrag
I denne oppgaven forsøkes det å identifisere generelle kategorier av utfordringer ved mapping av indekseringsspråk. I en todelt metode ble først dokumenter fra fire tidligere mappingprosjekter analysert og sju forskjellige kategorier av utfordringer ble identifisert. Fire av disse kategoriene ble viderebragt og eksemplifisert i en dataanalyse med et prekoordinert emneordssystem og en tesaurus som datagrunnlag. Analysen viser at utfordringer ved mapping av indekseringsspråk er til en viss grad mulig å generalisere. Utfordringene som ble identifisert ble knyttet til feilaktige mappinger, homonymi, hierarkisk plassering og forskjellig spesifisitet, sammensatte begreper, forskjellig praksis i bruk av emneordvokabularene og konteksten emneord står i, konsistens ved valg av mappinger og relasjonstyper og ressursbruk. Disse utfordringene var til stede i en forskjellig grad avhengig av om man mappet vokabularer basert på prekoordinering eller postkoordinering.

The goal of this thesis was to identify general categories of challenges in mapping of indexing vocabularies. Documentation from four previous mapping projects was analysed in order to identify common challenges. These were divided into seven categories. Four of these categories were then used in a data analysis using two new vocabularies, one consisting of compound subject headings and one thesaurus. The analysis shows that there are, to a certain degree, general categories of challenges in mapping of subject headings. The seven categories identified are connected to incorrect mappings, homonymy, hierarchy, compound concepts, different use of the vocabularies and the context of the subject headings, consistency in choice of mappings and the amount of resources needed for mapping. The extent of these challenges was depending on whether the mapping was based on mapping of pre-coordinated or post-coordinated vocabularies.

Høgskolen i Oslo og Akershus, Institutt for arkiv- bibliotek- og informasjonsfag
Oslo 2015

Forord

Å skrive masteroppgave har vært en langt mer fornøyelig prosess enn jeg hadde forventet meg. Ikke bare har det vært engasjerende og interessant, men jeg har også møtt stor velvilje og interesse over alt hvor jeg har henvendt meg innen kunnskapsorganiseringens kretser. Sånt motiverer enda mer, og noen takksigelser er på sin plass.

Takk til Kjersti Feiring Myrtrøen for interessante faglige diskusjoner og til Liv Bryn, Hege Nenseth og Ida Have for fremragende korrekturlesning, og takk til alle andre kollegaer ved Biblioteksentralen for støtte, interesse og engasjement. Takk til Eir Mariann Hvidsten og andre medstudenter for lange dager, store diskusjoner og gjensidig oppmuntring. Takk til Unni Knutsen og andre mappingengasjerte ved Universitetsbiblioteket i Oslo for imøtekommenhet, interesse, og tilbud om langt mer hjelp enn jeg har hatt vett til å benytte meg av. Takk til Elise Conradi og Ingebjørg Rype ved Nasjonalbiblioteket for tidlig interesse og gode råd. Takk til alle kjente som har fått svare på en rekke dumme spørsmål og bidra med sin fagkompetanse til begrepsavklaringer innen sykepleie, geologi, engelsk og andre fagområder. Til slutt en stor takk til veileder Ragnar Nordlie for god og kyndig veiledning.

Kristine Aalrust Kristoffersen

Oslo, 7. juni 2015

Innhold

Innledning

I takt med den teknologiske utviklingen har også kunnskapsorganisasjon fått et større spillerom og muligheter for å forbedre og utnytte data på nye måter. Datamaskinen gjorde det mulig å lagre og finne fram data på en mer effektiv måte enn man kunne med kortkatalogen. Internett har gjort det mulig å dele disse dataene med andre bibliotek, over hele verden om man vil. Det virker innlysende å gjøre det. Hvorfor skal vi gjøre samme jobben flere ganger, når det er mulig å utveksle og utnytte egen og andres arbeid og kompetanse? Den praktiske løsningen er selvfølgelig ikke så enkel. Foruten organisatoriske problemstillinger som opphavsrett og ansvarsfordeling, er datautvekslingen vanskelig å perfeksjonere fordi alle verdens mennesker, heldigvis, ikke tenker helt likt.

Når man tenker etter, er emneordsindeksering nesten for utrolig til å være sant. Ett menneske kan ta for seg et dokument, for eksempel en bok, og oppsummere hele denne bokas innhold i et par ord, som ikke en gang formuleres som setninger. Det er bare noen ord, kanskje tre eller fem, som til sammen sier noe om hva boka inneholder. Disse ordene registreres sammen med annen informasjon om boka i en database og så kan andre mennesker finne igjen denne boka basert på ordene. Ikke fordi de nødvendigvis vet at de leter etter nettopp denne boka, men fordi de formulerer det de ønsker seg med de samme ordene som boka har blitt beskrevet med. To store, uhåndgripelige størrelser, en boks innhold og et menneskes ønske om å lese noe med et bestemt innhold, har blitt kokt ned til de samme ordene, og disse ordene fungerer som innganger til riktig bok.

Et slikt utrolig sammentreff er mulig fordi vi mennesker tenker relativt likt. Vi oppfatter og sorterer virkeligheten og alt som finnes i det på omtrent samme måte, uten at noen nødvendigvis har bestemt det. De fleste setter søppelbøtta under kjøkkenvasken, legger bestikket i den øverste skuffen og buksene i den nederste hylla i klesskapet. I matbutikken ligger alle grønnsakene sammen, og kjøttet og fisken ved siden av hverandre. Vi tenker likt, men vi tenker ikke identisk. Vi er ikke datamaskiner. Vi kan se den samme filmen og etterpå diskutere hvorvidt det var en god film eller ikke fordi vår oppfatning av hva "god film" innebærer er forskjellig, og fordi vi ikke er enige i hvilke grep som gjør en film god. Dette er også indekseringens evige problem. Det er umulig å bli helt enige om nøyaktig hva et dokument handler om.

Emneord er, i motsetning til en filmtittel eller hvem som står for gitarspillet på en musikkplate, ikke noe man kan finne en fasit på. Man kan anta med en viss sikkerhet at mennesker i stor grad vil tolke et dokuments innhold likt og markere dette innholdet med lignende emneord. Men det er ikke alltid så lett. En sykepleier vil lete etter informasjon om en sykdom med et helt annet begrepsapparat og behov for detaljnivå enn en pårørende. Hvordan kan man få emneordene til å være nyttige for begge? Tradisjonelt har indekseringen vært samlingsspesifikk. Man har satt emneord til dokumenter i egen samling basert på behovet til brukerne i denne samlingen. Sykepleieren går til et medisinsk bibliotek, og de pårørende går på folkebiblioteket, og dokumentene om samme sykdom er indeksert med passende begreper i de respektive samlingene. Med utgangspunkt i de varierende indekseringsbehovene har man også gjennom årenes løp etablert forskjellige kontrollerte vokabularer for emneord, som utvikles og ivaretas etter hvert som gjenfinningsbehovet i de enkelte samlingene endrer seg.

Med den teknologiske utviklingen har man fått muligheten for å utveksle informasjon på tvers av disse samlingene. En av de vanligste tilnærmingene for denne utvekslingen, eller samordningen, er mapping. Mapping av emneord går, svært enkelt forklart, ut på å fortelle datamaskinen at to begreper som er benyttet for å beskrive dokumenter i to forskjellige samlinger på en eller annen måte er sammenlignbare. Gjennom en slik kobling, eller mapping, kan man øke tilfanget av emneord i vokabularene eller søke på tvers av samlinger indeksert med forskjellige vokabularer. I videre forstand betyr det at dokumentene som disse to ordene har beskrevet i de to samlingene er, på en eller annen måte, mulig å sette i relasjon til hverandre, for eksempel ved at de representerer det samme begrepet. Men er meningsinnholdet i et ord det samme i Norge og Finland? Eller for en sykepleier og en pårørende? Eller for to personer med samme jobb og utdanning som sitter rett ved siden av hverandre?

Siden man til en viss grad er enige om hvordan verden skal kategoriseres og sorteres, om hvilke ord som kan beskrive innholdet i et dokument, og hva disse ordene kan innebære, kan man si ja. Til en viss grad. Med utgangspunkt i denne enigheten er mapping mulig, og det åpner for nye muligheter for samordning. En mapping mellom to indekseringsspråk vil tilføre begge parter

mer informasjon og det vil gjøre det mulig å gjøre flere dokumenter gjenfinnbare ved hjelp av de samme emneordene.

Tanken bak mapping er altså god, men er det egentlig så lett? Enhver med kjennskap til klassifikasjon og indeksering vet at det ikke er "bare" å koke ned et helt dokuments innhold til fem ord, for ikke å snakke om fem ord som det kan tenkes at alle som måtte behøve å finne dokumentet også kommer på. For å gjøre det enklere for både indekserer og søker baserer man seg ofte på kontrollerte lister over emneord. Disse listene kan organiseres etter faste sett med regler for hvordan emneordene skal sorteres og struktureres, som for eksempel en tesaurus med forhåndsdefinerte interne relasjoner. Eller man kan benytte andre sett med regler, eller ingen spesiell sortering. Og så, når man ser at det gir mening, kan man finne på å lempe litt på reglene og etablere en intern praksis.

Å ta forskjellige valg som dette i indekseringsarbeidet er ikke feil. Det er å tilpasse seg det indekseringsbehovet man har, for å sørge for best mulig gjenfinning for brukerne man har. Sett i dette perspektivet ville det vært et dårlig valg å gjøre alt likt. Mapping handler ikke om å gjøre alt likt, men å utnytte disse forskjellene. Samtidig vil disse forskjellene også gjøre det vanskeligere å avgjøre hvilke emneord som skal mappes, nettopp fordi det er så mye mer knyttet til emneordene enn bare selve ordet.

Likevel er nytten større enn utfordringene, og dette er grunnen til at flere emneordssystemer, inkludert en rekke norske, er blitt mappet de siste årene. I disse dager pågår et samarbeid mellom Nasjonalbiblioteket (heretter NB) og Universitetsbiblioteket i Oslo (heretter UB) om å etablere en norsk generell tesaurus (heretter NGT). Forprosjektet består i to deler. En del ligger hos UB, og innebærer å undersøke metodikk for mapping av Humord mot DDC. Rapporten fra dette prosjektet beskrives nærmere i kapittel 2.6.1. Den andre delen av prosjektet innebærer å undersøke muligheter for og planlegge hvordan man skal etablere en slik tesaurus.

I rapporten fra denne delen av forprosjektet anbefales det at en generell norsk tesaurus skal bygges ut fra de mange vokabularene som allerede finnes hos NB og UB, med UBs tesaurus Humord som utgangspunkt. Prosjektet inviterte i startfasen våren 2014 et åpent seminar for

interesserte bibliotekarer og andre fra fagmiljøet hvor det grunnleggende ved en tesaurus ble diskutert (Ohren et al., 2015, s. 4-9). På dette seminaret ble også spiren til denne oppgaven sådd. Da seminaret fant sted var ikke fremgangsmåten hvor man med Humord som utgangspunkt eksisterende vokabular, slik at hvilken tilnærming som ville gi best resultat var oppe til diskusjon. Noen foreslo å oversette et eksisterende generelt vokabular, for eksempel amerikanske LCSH. Noen foreslo å lage et slags "lappeteppe" av eksisterende vokabularer. For meg virket det innlysende at det beste valget ville være å på en eller annen måte utnytte de allerede eksisterende og veletablerte vokabularene som fantes. Men hvordan? Og var det egentlig så enkelt som det hørtes ut, å bare koble forskjellige vokabularer sammen? En nærmere titt på hvordan-spørsmålet ble raskt besvart med "mapping." Men hvor enkelt, eller vanskelig, det egentlig var, var et mer komplisert spørsmål.

1.1 Motivasjon

For de fleste med innsikt i emneordvokabularers bruk og regler er det mulig å se for seg at et forsøk på å få to eller flere av disse til å passe sammen er en overkommelig, men utfordrende oppgave. Mange av utfordringene som forekommer er også lette å anta. Med en grunnleggende kjennskap til emneordsvokabularers struktur og ulikheter er det mulig å se for seg en rekke tilfeller av problemer som kan oppstå, med utgangspunkt i at de alle er organisert på en lignende måte. Likevel er alle emneordsvokabularer ulike, og det er nettopp dette som gjør det verdt å mappe dem.

Formålet med denne oppgaven er å undersøke om det er mulig å identifisere noen fellesnevnere for de utfordringene som oppstår i forskjellige forsøk på mapping hvor det er brukt ulike tilnærminger til mapping av indekseringsspråk med forskjellig struktur. Deretter undersøkes det om disse utfordringene igjen oppstår med to nye sett med emneord. Man kan også si at det forsøksvis undersøkes om emneordvokabularer er ulike på en lik måte.

En identifisering av slike utfordringer vil forhåpentligvis være nyttig ved oppstart av nye mappingprosjekter, et slags frampek som man kan ta hensyn til i valg av mappingmetoder og indekseringsspråk som skal mappes. Samtidig sier utfordringer i mapping noe om svært sentral tematikk rundt kunnskapsorganisasjon i dag: menneskeskapte systemer forsøkes å organiseres på en måte som også kan utnyttes på best mulig måte av datamaskiner. Alle de eksisterende

mappingprosjektene ville ikke blitt gjennomført med mindre teknologien lå til rette for det. De fleste av utfordringene som oppstår kan sies å oppstå fordi det ligger en menneskelig bevissthet, både individuell og felles, til grunn i alle indekseringsspråk. De bygger på vår felles forståelse og enighet, i den grad man kan være enige, om hvordan kunnskap skal organiseres og sorteres.

1.2 Problemstilling

Det er ikke er noen garanti for at to forskjellige mennesker benytter samme emneord nøyaktig likt (snarere vet vi at det ikke er tilfelle), men vi vet også at mennesker tenker *nesten* likt. Vi assosierer nesten de samme tingene med ordene. Vi har en felles enighet om hvilke ting som hører sammen og hvilke måter disse tingene kan sorteres på. Med utgangspunkt i denne kombinasjonen av enighet og uenighet baserer denne oppgavens problemstilling seg på en antagelse: de samme typene utfordringer vil oppstå, i større eller mindre grad, i de fleste mappinger av indekseringssystemer. Uavhengig om emneordene er basert på pre- eller postkoordinering, hvorvidt de benyttes av fag- eller folkebibliotek, hvorvidt det mappes mellom to emneordsvokabular eller fra et emneordsvokabular til et klassifikasjonssystem, vil enigheten og uenigheten i hvordan kunnskap kan organiseres føre til liknende utfordringer. Dette leder til følgende problemstilling:

Hva slags utfordringer kan oppstå ved forskjellige mappinger av indekseringsspråk?

- Hvilke fremgangsmåter er benyttet og hvilke utfordringer har oppstått i tidligere prosjekter?
- Hvilke fordeler og ulemper innebærer de forskjellige fremgangsmåtene?
- Hvilke fordeler og ulemper innebærer mapping av forskjellige typer indekseringsspråk?

Vurderinger, begrepsoppklaring og tanker rundt problemstillingen diskuteres nærmere i oppgavens metodekapittel.

1.3 Oppgavens oppbygning

I kapittel 2, teori og tidligere forskning, presenteres først standarden for interoperabilitet mellom emneordssystemer samtidig som begreper og teknikker rundt mapping avklares. Deretter presenteres noen grunnleggende prinsipper for emneordsvokabularer og bruk av disse. Til slutt

presenteres relevant forskning og et utvalg mappingprosjekter som ansees som spesielt relevante eller interessante for oppgaven.

Kapittel 3 er oppgavens metodekapittel hvor alle metodiske valg og avveininger presenteres. Her presenteres også problemstillingen og forskningsspørsmålene i detalj, og noen begreper avklares. Begrensninger og valg av dokumenter og data til analysen beskrives også her.

Kapittel 4 inneholder en detaljert beskrivelse av de fire prosjektene som er grunnlaget for dokumentanalysen. Her omtales prosjektenes datagrunnlag, valg av mappingteknikker og også hva som ble ønsket oppnådd med prosjektene. Dette kapittelet er ment for å gi bedre oversikt og forståelse av prosjektene i forkant av dokumentanalysen.

Dokumentanalysen og funn knyttet til denne beskrives i kapittel 5. Kapittelet er sortert etter de sju kategoriene med utfordringer som ble identifisert, samt en oppsummering. I kapittel 6 undersøkes det hvorvidt kategoriene i dataanalysen er av en allmenn karakter. Dette gjøres ved å eksemplifisere kategoriene med data fra to nye vokabularer. Kapittelet oppsummeres med noen overordnede inntrykk.

I kapittel 7 diskuteres funnene i analysens to deler og funnenes samsvar med tidligere forskning fra teorikapittelet. I kapittel 8 oppsummeres funnene og diskusjonen, og det gjøres noen tanker om mapping og forskjellige utgangspunkt for mapping. Det gjøres også noen tanker om og forslag til videre forskning.

2 Litteratur

I dette kapittelet presenteres en gjennomgang av litteratur som er vurdert som relevant for oppgaven. Innledningsvis presenteres den internasjonale standarden for interoperabilitet mellom tesaurus og andre emneordssystemer. I den forbindelse avklares mappingbegrepet og de mest sentrale prinsippene som ligger til grunn for mapping av emneordsystemer i dag. Videre omtales grunnleggende prinsipper for indekseringsspråk med eksempler i Biblioteksentralens emneord, Humord og Deweys desimalklassifikasjon. Deretter problematiseres forskjellen på prinsippene i disse indekseringsspråkene. Omfanget av forskning med samme fokus som i denne oppgaven er svært begrenset. Litteraturen på området er preget av muligheter for fremgangsmåter ved mapping og oppsummeringer av konkrete prosjekter.

Det er, etter det jeg kjenner til, viet lite oppmerksomhet til å sammenligne erfaringer fra mappingprosjekter som er gjennomført eller hvorvidt man kan uttrykke noen generelle former for utfordringer ved mapping. Likevel er noe arbeid gjort, og dette presenteres her sammen med en oppsummering av utviklingen av mappingrelasjoner før den nåværende standarden. Avslutningsvis presenteres tre prosjekter som eksemplifiserer hvilke nyttefunksjoner man kan ha ved mapping og som har problematisert noen fenomener ved mapping på en generell basis.

2.1 ISO-standard 25964-2:2013

Den gjeldende standarden som benyttes i forbindelse med mapping av emneordsvokabularer er ISO 25964, og spesielt del 2, utgitt i 2013. Det betyr at en rekke mappinger, blant annet MACS og CrissCross som omtales senere, ble utført og fullført før denne standarden ble utgitt. Tidligere prosjekter er likevel utført med liknende formål og med liknende metoder, slik at sammenligningsgrunnlaget er godt. Det er denne standarden som legges til grunn for begrep og definisjoner i denne oppgaven.

Standarden definerer mapping både som verb og substantiv. Prosessen mapping (å mappe) defineres som "prosessen å etablere relasjoner mellom begrepene i ett vokabular med begrepene i et annet" og substantivet mapping defineres som resultatet av prosessen: "relasjonen mellom et begrep i et vokabular og ett eller flere begrep i det andre." (ISO 25964-2, 2013, s. 7, mine oversettelser).

2.1.1 Modeller for mapping

Standarden beskriver flere modeller for mapping - fremgangsmåter for å mappe et vokabular til et annet. En tilnærming er direkte mapping. Man mapper et begrep i ett vokabular direkte til et begrep i et annet vokabular, eller, dersom det skal mappes mellom flere vokabularer, mapper alle vokabularer direkte til hverandre. En annen modell kalles "hub structure" og innebærer at alle vokabularer som skal mappes sammen mappes mot en felles "hub". På norsk blir dette ofte kalt nav. Et vanlig nav i denne modellen er Dewey i forskjellige oversettelser.

En tredje modell er selektiv mapping, hvor man mapper bare deler av vokabularene. Dette kan gjøres ved et lite overlapp av sammenfallende fagområder innen de to vokabularene. En annen modell for selektiv mapping er å velge mappinger basert på hvilke emneord som er knyttet til dokumenter i katalogen.

Standarden anbefaler en direkte mapping for tesauruser som skal benyttes på flere språk, mens mapping mot et nav og selektiv mapping anbefales for vokabularer som er utviklet og benyttet hver for seg. Det skilles også mellom retningen på mappingene (ISO 25964-2, s. 17-20). Selv om det er mulig å mappe begge veier mellom to vokabularer, er det vanligste å velge et kildevokabular og et målvokabular. Man mapper fra et kildevokabular og til målvokabularet. Dette åpner for flere typer relasjoner, som ikke trenger å kunne reverseres. For eksempel mapper man fra et kildevokabular til målvokabularet Dewey når man mapper til Dewey som nav.

2.1.2 Mappingtyper

Standarden opererer med flere typer mappingrelasjoner. Disse benyttes for å uttrykke forskjellige forhold mellom emneordene som mappes. I rapporten "Metodikk for mapping av Humord mot WebDewey" fra Universitetsbiblioteket i Oslo oppsummeres de slik:

=EQ	Likhetstegnet angir at mappingen er eksakt
~EQ	Tilden angir at mappingen er ikke eksakt. Det innebærer at begrepene kan være like i noen sammenhenger, men ikke i alle eller at begrepene kan være delvis overlappende eller avvike noe i betydningsinnhold
BM	Broader mapping. Termen i målvokabularet har en videre betydning enn termen i kildevokabularet
NM	Narrower mapping. Termen i målvokabularer har en mer spesifikk betydning enn termen i kildevokabularet
RM	Related mapping. Termen i målvokabularet assosieres med termen i kildevokabularet, men er ikke et synonym, et kvasisynonym eller en bredere eller smalere term

(Gulbrandsen, Heggø, Knutsen & Seland, 2015, s. 8)

=EQ uttrykker med andre ord at to emneord beskriver det samme begrepet. Det kan være identiske ord eller synonymer. Denne relasjonen kan gå begge veier. ~EQ beskriver en relasjon mellom to nesten like emneord. BM og NM kan benyttes spesielt ved mapping av vokabularer med en hierarkisk struktur. RM uttrykker det som kalles en assosiativ relasjon, gjerne de samme forbindelsene som uttrykkes med en se også-henvisning i en tesaurus.

Standarden åpner også for en mapping som ivaretar sammensatte begreper. "Compound equivalence" (sammensatt ekvivalens) kan benyttes når det mappes fra et sammensatt begrep i kildevokabularet, og begrepet er splittet i to begrep i målvokabularet. Det er igjen to typer sammensatt ekvivalens. "Intersecting compound equivalence" uttrykkes EQ+, og fungerer som den boolske operatoren AND. For eksempel kan begrepet "women executives" mappes til Women og Executives. Denne typen mapping er ikke reversibel.

Den andre typen er "cumulative compound equivalence", som uttrykkes EQ|. Denne mappingen kan minne om den boolske operatoren OR. Likevel er det ikke helt likt. Dersom summen av to begreper i målvokabularet kan representere et sammensatt begrep i kildevokabularet kan dette uttrykkes med EQ|. Eksempler som gis er at "Inland waterways" kan uttrykkes med Rivers og Canals, eller at "Hoisery" kan uttrykkes med Stockings og Socks (ISO 25964-2, 2013, s. 22-24).

2.1.3 Mapping av klassifikasjonssystemer

Et klassifikasjonssystem og et emneordssystem har en forskjellig tilnærming til sortering av emnene. Klassifikasjonssystem sorteres etter fag, og ikke emne (Hjortsæter, 2009, s. 41). Dette innebærer at ett emne kan ha flere plasseringer, avhengig av hvilket fag det faller innefor. For eksempel kan emnet Hest behandles rent biologisk, men også innen husdyrhold, gårdsdrift, jakt, transport og innen idrett, for å nevne noe. I tillegg har man hest som objekt i for eksempel kunst og hest som produkt i for eksempel matlaging. Denne grunnleggende forskjellen mellom indekseringsspråkene spiller en viktig rolle når man skal mappe mellom et emneordssystem og et klassifikasjonsskjema. Spesielt i postkoordinerte emneordsvokabularer vil emner bare ha en plassering i hierarkiet, og så vil sammensatte begreper uttrykkes ved kombinasjon med emner fra andre steder i hierarkiet. Til tross for denne forskjellen er mapping fra emneordsvokabularer til klassifikasjonsskjemaer, blant annet Dewey, svært vanlig. Standarden gir en rekke anbefalinger for mapping mellom tesaurus og klassifikasjonsskjema.

For det første påpekes det at man må vurdere en klasses omfang basert på flere kilder, blant annet ved å se på beskrivelser knyttet til klassen og klassens hierarkiske plassering. Man kan mappe begge veier, og i noen tilfeller med relativt enkle begreper vil det være mulig å mappe ett begrep fra en tesaurus til ett begrep i et klassifikasjonsskjema. De fleste klassifikasjonsskjema åpner for nummerbygging, noe som innebærer at de fleste mappinger vil måttte være sammensatte ekvivalensmappinger. Standarden anbefaler også å mappe til klassebetegnelsene, ikke til beskrivelser og ord tilknyttet klassebetegnelsene (ISO 25964-2, 2013, s. 55-56).

2.1.4 Mapping av prekoordinerte begreper

Prekoordinerte begreper kan være flere ting, både emneord i streng og bygde numre fra et klassifikasjonsskjema. Felles for dem er at et sammensatt begrep uttrykkes ved å knytte flere

elementer sammen. Denne sammensetningen gjøres i selve vokabularet og er dermed en permanent kombinasjon.

I forbindelse med prekoordinerte begreper slår standarden fast at alle typene mapping kan benyttes, dersom det finnes en direkte match i det andre systemet. Her opprettes en en-til-en-mapping. Hvis et prekoordinert begrep passer med flere begreper i målvokabularet er det mulig å opprette en en-til-mange-mapping. Man kan for eksempel splitte et bygget nummer og mappe hver av bestanddelene i klassenummeret til passende begreper i målvokabularet. Ved mapping fra prekoordinerte begreper til en tesaurus advarer standarden mot at mappingen baseres på hvilke emneord som er blitt postkoordinert i et dokument. Kombinasjonen av to begreper kan uttrykke mange forskjellige forhold mellom de to, mens et klassenummer gjerne bare uttrykker ett begrep (s. 32-35).

Når det gjelder mapping fra vokabularer basert på emneord i streng mot en tesaurus som målvokabular foreslår standarden tre tilnærminger. Den første er ganske enkelt å splitte alle strengene og mappe hvert ord til målvokabularet. Denne tilnærmingen gjør det mulig å se bort fra de, til tider, kompliserte reglene for prekoordinering av emneord, men fjerner også muligheten for å uttrykke mer komplekse begreper. Dette kan føre til lavere presisjon ved søk. Å ivareta sammensatte begrep er derimot mulig i løsning nummer to, hvor man i tillegg til å gjennomføre den første løsningen mapper alle strenger til tesaurusen. Dersom et sammensatt begrep er uttrykt i både kildevokabularet og målvokabularet, mappes dette direkte. Dersom et sammensatt uttrykk kan mappes ved hjelp av sammensatt ekvivalensmapping, gjøres dette.

Disse fremgangsmåtene krever trolig en form for intellektuell kontroll for å unngå forvirring og feilkoblinger. Den tredje metoden som foreslås er å kun mappe strengene som har blitt benyttet til å indeksere dokumenter i en samling. Dette forbedrer både presisjon og fullstendighet, men har igjen en bakdel ved at mappingene bare er garantert å være aktuelle for den dokumentsamlingen som er blitt benyttet som grunnlag (s. 69-71).

2.2 Indekseringsspråk

Emneord er den beste søkeinngangen for folk som ikke vet hvilken bok de skal ha. De vet bare hva de skal ha svar på, eller hva de føler at de burde ha svar på, eller hva de tror at de må ha svar på. Dette behovet for svar kan forsøkes formulert som et enkelt ord, som igjen kan hente fram alle dokumenter som har blitt tilordnet dette ordet fordi de antas å ha noe med dette ordet å gjøre. Ganske enkelt og ganske vanskelig. For hva handler egentlig en bok om? Og hvilke ord uttrykker hva denne boka handler om? Gjennom tidene har det blitt etablert flere måter å organisere ordene vi benytter for å uttrykke noe om et dokuments innhold. Det kan ganske enkelt være en liste av ord, det kan være en liste med ord sortert i grupper, det kan være at listen også inneholder noen synonymer man ikke bruker, eller noen andre ord som man kanskje kan få bruk for i forbindelse med det første ordet, og så videre.

I emneordsindeksering skiller man mellom to måter å kombinere emneord for å uttrykke et dokuments sammensatte innhold på: postkoordinering og prekoordinering. Mens prekoordinering av emneord uttrykker sammensatte begreper ved å kombinere emneordene i vokabularet, kombineres emneordene ved postkoordinering i tilknytning til hvert enkelt dokument som indekseres, og de sammensatte begrepene må dermed også uttrykkes gjennom flere søkeord ved søk.

2.2.1 Postkoordinering og tesaurus

Emneordsvokabularer basert på postkoordinering består av enkeltstående ord, som kan fritt kombineres ved indeksering for å uttrykke sammensatte begreper og et dokuments innhold. En slik måte å indeksere på ansees som mer fleksibel og har blitt mulig i en større grad gjennom den teknologiske utviklingen, til forskjell fra tiden hvor man var prisgitt kortkatalogens plassbegrensninger. Emneord basert på postkoordinering kan organiseres på flere måter, for eksempel en ren alfabetisk liste med flat struktur, eller som en tesaurus.

En av de mest regelbundne formene for organisering av emneord er en tesaurus. Et fast sett med henvisninger uttrykker hver terms forhold til andre termer i vokabularet. I tillegg til hierarkisk plassering, det vil si hva som er overordnet, underordnet og likestilt en term, uttrykkes synonymer som ikke brukes med en se-henvisning, og andre relevante termer uttrykkes med en

se også-henvisning. Se også-henvisningene har igjen regler for hva slags relasjoner det skal være mellom ordene for at denne henvisingen skal benyttes. I tillegg til dette har en tesaurus oppdeling i fasetter.

Humord er en tesaurus som i 2011 inneholdt nærmere 25 000 termer innen humaniora og samfunnsfag med tilgrensende områder. Emneordene består i innholdsbeskrivende emneord og formtermer. Hierarkiene i Humord er ordnet etter fagområder (Hougaard, 2011, 11. juli.) Som andre tesauruser har Humord en hierarkisk ordning, med overordnede og underordnede termer i flere nivåer.

2.2.2 Prekoordinerte emneord

Prekoordinerte emneord uttrykker også sammensatte begreper, men her ligger kombinasjonen "lagret" i selve vokabularet. Den vanligste måten å prekoordinere emenord på er å sortere emneordene i strenger, etter bestemte regler om rekkefølge og struktur. Senere i oppgaven omtales ofte prekoordinerte emneord som emneord i streng. En slik organisering av emneord gjør det mulig for både indekserer og sluttbruker å lettere finne frem til sammensatte begreper uten å måtte formulere og kombinere selv. Enkeltstående emneord kan også inngå i et emneordssystem basert på prekoordinerte emneord, dersom de alene uttrykker et begrep godt nok. Samtidig ansees prekoordinerte emneord som vanskeligere å ivareta og etablere på en konsistent måte. Det er svært vanlig å sette flere emnestrenger på ett dokument, altså å postkoordinere prekoordinerte emneord.

Et av de største prekoordinerte emneordsvokabularene i Norge i dag er Biblioteksentralens emneord. Biblioteksentralen eies av landets kommuner og fylkeskommuner og leverer bøker og tjenester til bibliotekene. Dette inkluderer også, for de aller fleste folkebibliotek og endel skolebibliotek, katalogposter med emneord. Både skjønn- og faglitteratur får emneord.

Biblioteksentralens emneordsystem (heretter omtalt som BIBBI) er et prekoordinert emneordssystem som pr. mai 2015 består av nærmere 40 000 emnestrenger. Disse er fordelt over flere typer emneord: Generelle emneord, korporasjoner som emner, personer som emner, konferanser, standardtitler og geografiske emneord. Emnestrengene benyttes av de fleste folkebibliotekene i landet, som utgjør størsteparten av Biblioteksentralens kunder. Knyttet til

hver emnestreng er det et deweynummer. Dokumenter fra 2000 og nyere har emneord tilknyttet numre fra DDK5, eldre dokumenter er indeksert etter DDK4. I tillegg er 15 000 emnestrenger fra de generelle og de geografiske emneordene tilknyttet numre fra den nye oversettelsen av Dewey som lanseres høsten 2015 (Kjersti Feiring Myrtrøen, e-post, 7. og 12. mai, 2015). Emnestrengene konstrueres etter retningslinjer angitt i Hjortsæters " Emneordskatalogisering. Innholdsanalyse, emnerepresentasjon og lagring" fra 2009. Dette innebærer blant annet at ordet som ansees som viktigst settes først i strengen.

Prekoordinerte emneords styrke og svakhet er muligheten til å uttrykke sammensatte begreper i emneordsvokabularet. På denne måten er det mulig å uttrykke et dokuments innhold i en enkelt emnestreng. Ulempen med strenger er at de er nettopp kompliserte, og lages etter regler som kan være tidkrevende å følge. Dessuten er meningen som legges i ordenes rekkefølge ikke alltid lett for en datamaskin å lese. Da Nasjonalbiblioteket evaluerte sin emneordspraksis for noen år siden, ble det besluttet å gå bort fra prekoordinerte emneord, til tross for at flere store systemer som Library of Congress Subject Headings (LCSH) og Svenska Ämnesord (SÄO) er prekoordinerte:

"De siste årene har det vokst fram en erkjennelse av at prekoordinerte emnesystemer er svært vanskelig å holde ved like, og dessuten er lite brukervennlige. Ikke minst har LCSH vært gjenstand for slik kritikk. Reglene for sammensetninger er komplekse, setter (for) store krav til indekserer, noe som i sin tur fører til mye feil. Tilsvarende erfaring er gjort med SÄO. Den noe større uttrykkskraften man oppnår ved å tillegge rekkefølgen av emneordene mening, blir i mange tilfeller svært kostbar. Prekoordinerte emneord er heller ikke godt tilpasset søkemotorer, og slett ikke semantisk web. " (Ohren, Rydland & Rype, 2013, s. 16).
Å organisere prekoordinerte emneord for semantisk web er relativt upløyd mark. Kanskje nettopp fordi emneordenes rekkefølge i en streng ikke kan tolkes av en maskin uten videre.

2.2.3 Deweys desimalklassifikasjon

Det finnes mange klassifikasjonssystemer, men Dewey kan sies å være et av de mest etablerte og utbredte internasjonalt. Det er også et av de mest sentrale i forbindelse med mapping. En rekke vokabularer på flere språk er i de siste årene blitt mappet til Dewey. Selv om Dewey eksisterer i mange oversettelser, er klassifikasjonsnumrene en internasjonal fellesnevner uavhengig av språk.

Selv om variasjoner i innhold og tolkning så klart forekommer er det en rimelig antagelse at samme deweynummer vil benyttes på dokumenter med mange fellesnevnere uavhengig av hvor i verden dokumentet indekseres.

Likevel er det interessant at nettopp Dewey er blitt så sentralt i mappingarbeid da det er et klassifikasjonssystem og ikke et emneordsvokabular. Dewey er organisert og strukturert på helt andre måter enn et emneordsvokabular. For det første er det, som tidligere diskutert om klassifikasjonsskjemaer generelt, sortert etter fag framfor emne. I tillegg åpner Dewey for nummerbygging for å uttrykke sammensatte begreper, altså en form for prekoordinering.. Likevel er mapping mot Dewey som nav blitt svært vanlig å gjennomføre, og flere eksempler på dette blir omtalt i oppgaven. En av de viktigste årsakene til mappingen mot Dewey kan være de mange oversettelsene av Dewey som eksisterer. Blant annet er en norsk oversettelse av Dewey, kalt WebDewey, planlagt ferdigstilt høsten 2015.

Merk at Deweys desimalklassifikasjon omtales på flere måter i denne oppgaven. Både som Dewey og DDC, som et betegnelse på systemet generelt, uavhengig av språk og versjon, og som WebDewey, som en betegnelse på den nyeste oversettelsen av Dewey på norsk, og som DDK5, den inntil videre benyttede versjonen av Dewey på norsk.

2.3 Indekseringskonsistens og emneord i kontekst

Når man graver seg ned i emneordvokabularenes struktur og finurligheter er det lett å glemme at de faktisk har en svært praktisk funksjon. De er indekseringsverktøy, og hele systemet er opprettet og vedlikeholdt med et formål. Emneordsvokabularet skal sikre en god og konsistent indeksering av dokumentene i samlingen. Her menes samlingen i videste forstand - alle dokumentene som er indeksert ved hjelp av vokabularet, uavhengig av hvor mange institusjoner som benytter vokabularet.

Man kan si at emneordsvokabularet og samlingen er i en slags symbiose. Man oppretter sjelden nye emneord uten at et dokuments innhold ikke kan beskrives med eksisterende emneord. Dermed vokser emneordsvokabularets univers i takt med samlingens kollektive innhold. Unntaksvis kan man nevne de største systemene som for eksempel Dewey, som benyttes uten at

hvert bibliotek som bruker det har mulighet til å innføre nye begreper uten videre. Men i de fleste indekseringssystemer er utviklingen av systemet og tilveksten i samlingen tett knyttet sammen. I indekseringsarbeidet er det et svært vanlig grep å, når man tror man har funnet rett emneord, gå inn i samlingen og undersøke hvilke andre titler som allerede er tilknyttet dette emneordet.

På denne måten tilfører hvert tilfelle av indeksering med et emneord ny mening til emneordet i form av "dette emneordet beskriver dette dokumentet." Når flere dokumenter er knyttet til samme emneord blir summen av disse dokumentene (eller snarere summen av katalogpostene til disse dokumentene) noe som tilfører emneordet mening. Det har også betydning hva hver indekserer vurderer som relevant for sine brukere, hvor mye tid og interesse indekserer vier dokumentet, og så videre. Disse vurderingene har en innvirkning på det som kalles indekseringskonsistens. I hvor stor grad er mennesker i stand til å vurdere et dokuments innhold likt og representere dette innholdet på samme måte? Naturligvis er ikke mennesker i stand til å gjøre identiske vurderinger.

Det er nettopp på grunn av denne umuligheten i å oppnå en perfekt indekseringskonsistens man vier tid og ressurser til å vedlikeholde kontrollerte emneord, og det er også en av grunnene til å sette i gang med et stort og omfattende mappingarbeid - for å oppnå en indekseringskonsistens med samlinger indeksert med andre vokabularer. Samtidig tilfører denne inkonsistensen emneordene mening. Flere vurderinger av passende bruk av et emneord gir flere tolkninger av emneordet. Flere vurderinger av passende emneord for et dokument gir flere innganger til dokumentet.

Dette forholdet mellom språkbruk, språkbrukere og konteksten språket brukes i kalles i språkvitenskapen pragmatikk. Pragmatikken studerer kontekstens bidrag til ytringens mening (Vagle, Sandvik & Svennevig, 1993, s. 20). Selv om pragmatikken som regel forholder seg til en litt mer omfangsrik tekst enn enkeltstående ord er det likevel relevant å se på det i forbindelse med mapping, nettopp fordi emneord får tilført så mye mening fra konteksten de oppstår i. Med emneord forsøker man i praksis å oppsummere en hel bok, film eller artikkels innhold med et lite knippe ord. Det er klart at man da legger mye mening i emneordene. Og denne meningen er avhengig av konteksten.

Gödert, Hubrich & Nagelschmidt (2014) skiller mellom "core meaning" og "conceptual meaning" i termer. "In their core meaning, entities refer to concrete or abstract, real-life or fictive things in the world. Their core meaning is independent of the information resources to which the entities are assigned. In their contextual meaning, entities refer to aspects of concrete or abstract, real-life or fictive things in the world, putting things in relation to other. Their contextual meaning can be dependent of the information resources to which the entities terms are assigned." (s. 108). Dette skillet tar hensyn til det som pragmatikken kalles "kontekstens bidrag til ytringes mening," hvor man kan tolke termers mening både i og utenfor konteksten de står i, altså dokumentene de er blitt knyttet til. I tillegg spiller som nevnt faktorer som termenes plassering i vokabularets struktur, forklaringer som er knyttet til dem og annen informasjon også inn i konteksten.

Med utgangspunkt i disse typene mening hos termer skiller Gödert, Hubrich & Nagelschmidt også mellom "focused mapping" og "comprehensive mapping." I focused mapping tar man bare hensyn til "core meaning," mens man i "comprehensive mapping" tar hensyn til begge typer mening (s. 108-110). En fullstendig automatisk mapping basert på tekstlikhet mellom emneordene vil for eksempel være en focused mapping. En focused mapping vil generere langt færre mappinger enn en comprehensive mapping. Selv om man skiller tydelig i hvor stor grad man tar hensyn til termenes kontekst, kan man likevel ikke inkludere all mening: "[...] we must accept that every real world concept has by definition its own meaning which cannot be shared by any other concept identically. An understanding of true conceptual exchangeability can therefore not be derived from real world observations but only within the special settings of benefits expected of task-oriented tools." (s. 237).

2.4 Tidligere relasjonstyper

Ettersom standarden som nå foreligger er fra 2013 er mesteparten av mappingarbeidet som er gjort foretatt før publiseringen av denne. Dermed er også mappingtypene som standarden etablerer ikke de som har blitt benyttet i alle prosjekter. Mappingtypene i standarden er i aller høyeste grad preget av relasjonene som finnes innbyrdes mellom termer i en tesaurus. Dette er et naturlig valg ettersom standarden foreslår tiltak for interoperabilitet mellom en tesaurus og andre emneordsvokabularer. Sånn sett finnes det ingen standard for interoperabilitet mellom andre

typer emneordsvokabularer hvor ingen av partene er en tesaurus, men mange av prinsippene er så klart de samme. For eksempel vil mappinger for fullstendig likhet og nesten fullstendig likhet være relevant uansett hva slags vokabularer man mapper.

Allerede i 1995 foreslo Margaret A. Chaplan forskjellige typer mapping, basert på hva slags relasjoner som ble benyttet i en mapping fra Laborline Thesaurus til Library of Congress Subject Headings. Her ble det benyttet hele 19 forskjellige kategorier av relasjoner mellom termene. De spenner fra fullstendig likhet til ingen likhet. I tillegg er det egne kategorier for for eksempel forskjellige stavemåter, hvorvidt emneordet er oppgitt i entall eller flertall, hvorvidt ordene i termen er oppgitt i forskjellig rekkefølge, og mange flere små variasjoner. Noen kategorier oppgir også tilfeller hvor man kan mappe til et synonym i tesarusen eller underordnet og overordnet term. Mye av dette er det samme som standardens mappingtyper sier noe om. Men de 19 kategoriene inneholder også mer spesifikke relasjoner. En kategori gjør det mulig å mappe termer som motsetninger av hverandre. En annen benyttes dersom termen er oversatt, for eksempel hvis et latinsk navn er benyttet i det ene vokabularet og et folkelig begrep er benyttet i det andre. Polysemer har en egen kategori (Chaplan, 1995).

Mange av disse kategoriene er inkludert i standardens mappingtyper. Mappingtypene i standarden favner videre enn Chaplans kategorier. Små variasjoner i stavemåte eller rekkefølgen på ord gjør for eksempel ikke mappingene ukvalifisert for en EQ eller ~EQ-mapping. Samtidig sier ikke en nesten eksakt match noe om hva som gjør at det bare er nesten likt. I etterkant av Chaplans kategorier ble det foreslått mange andre forskjellige mappingtyper, og i 2008 tok McCulloch & MacGregor tak i Chaplans 19 kategorier og undersøkte hvorvidt de var nøvendige. Det ble undersøkt hvorvidt et så stort antall typer mapping var unødvendig stort og om de kunne kortes ned. For å undersøke dette ble de 19 kategoriene benyttet av flere personer til mapping av flere ulike vokabularer, og deretter ble valget av mappingtyper sammenlignet. Resultatene viste at selv om enigheten i valg av mappingtyper var på 82%, var ikke på langt nær alle de 19 kategoriene nødvendige.

Bare ni kategorier ble validert gjennom forsøkene. Basert på bruken av disse ni ble det slått fast at fremtidige mappingtyper måtte kunne uttrykke eksakt match, bredere eller smalere match og

"Concept match." Interessant nok var kategori 14, concept match, den kategorien det var desidert størst utenighet om. Kategori 14 betegner en mapping mellom like begreper som er betegnet med forskjellige emneord i vokabularene uten at ordene er oppgitt som synonymer. Det foreslås at selv de ni validerte kategoriene kan kuttes ned ytterligere. Avslutningsvis foreslås det noen prinsipper for valg av mappingtyper. Mappingene som benyttes bør være mulig å benytte på flere typer kunnskapsorganisasjonssystemer, tilpasset semantisk web, og kunne opprettholde en høy grad av spesifisitet uten å være vanskelig å bruke. De bør også være godt definert for å unngå forvirring angående bruken (McChulloch & MacGregor, 2008). Disse typene mapping kan man kjenne igjen i standardens mappingtyper i dag.

I CrissCross-prosjektet, som omtales nærmere i kapittel 4, ble det benyttet et sett med relasjonstyper som heller fokuserer på hvor stor likhet det er mellom begrepene framfor å betegne karakteren på relasjonen mellom dem. CrissCross ble avsluttet i 2010, før ISO-standarden ble utgitt, og bruker derfor andre typer mappinger enn dem som er anbefalt i standarden. Likevel ligner de ganske mye. "Degrees of Determinacy" ble laget for mapping av tyske SWD (Schlagwortnormdatei) til DDC og er derfor beregnet spesielt for mapping mot Dewey (Gödert, Hubrich & Nagelschmidt, 2014, s. 139).

Mappingene beskrives med fire forskjellige "Degrees of Determinacy" (likhetsgrader) D4 er den største graden av likhet, og tilsvarer det som i standarden betegnes som =EQ, eksakt samsvar. D3 er "a slight degradation of D4", og er i stor grad tilsvarende ~EQ, tilnærmet eksakt samsvar. D2 brukes ved mindre samsvar, for eksempel når en term passer innunder et klassenummer sammen med en rekke andre termer. Dette minner noe om BM (Broader mapping) i standarden. D1 er den minste graden av samsvar og betegner "slight conceptual congruency", som for eksempel assosiative relasjoner (Jacobs, Mengel & Müller, 2010, s. 39-40).

2.5 Tidligere forskning

Som nevnt er det gjort få sammenligninger av mappingprosjekter. Selv om mappingprosjektene er mange, og de alle viser til hverandre, er rapporter og artikler knyttet til prosjektene ikke spesielt fokusert på erfaringer fra disse prosjektene. Dette betyr forøvrig ikke at man ikke er klar over dem før man setter i gang med egne prosjekter, men det er viet lite oppmerksomhet til

sammenligning av erfaringer i litteraturen. En mulig årsak til at det er viet lite oppmerksomhet til å sammenligne mappingprosjekter kan være at man, med en innsikt i emneordsvokabularers struktur og forskjellige særegenheter, enkelt kan gjøre svært gode antagelser om hvilke utfordringer som kan oppstå. Det er dessuten en mulighet for at man setter i gang med mappingprosjekter med en innstilling om at ja - det kommer til å være ressurskrevende, og ja - det kommer til å være problemer underveis. Disse antagelsene bygger på samme antagelse som resten av oppgaven: emneordsvokabularer ligner på hverandre, men de er ikke like.

Mangelen på litteratur om temaet kan være et tegn på at det nettopp finnes allmenne typer utfordringer, uavhengig av hvilke vokabularer man mapper, og ingen skriver om det fordi "alle" skjønner at disse utfordringene vil oppstå. I denne oppgaven forsøkes det å definere hva disse utfordringene innebærer, hvorvidt det finnes fellestrekk mellom utfordringene, for så å forsøke å gjenskape utfordringene i et nytt sett med vokabularer.

En kort oppsummering av fordeler og potensielle utfordringer ble publisert i 2005, hvor McChulloch og Nicholson blant annet påpeker at en av de største ulempene ved mapping er at det er svært arbeidskrevende og at man behøver en viss grad av intellektuell kontroll. Oppsummert fra tidligere prosjekter fremheves også forskjellige grader av spesifisitet og mapping mellom begreper med forskjellig hierarkisk plassering som utfordrende. I forbindelse med mapping til LCSH har det også vært vanskeligheter knyttet til prekoordinering. Både sammensatte begreper og underinndelinger i strengene fremheves her, samt forskjellige tilnærminger til å håndtere for eksempel synonymer og homonymer (McCulloch & Nicholson, 2005).

I Gödert, Hubrich & Nagelschmidt diskuteres problemer ved å etablere semantisk interoperabilitet mellom indekseringsspråk. Som et utgangspunkt gir de dette eksempelet:

Thesaurus 1	Thesaurus 2
Library BF Public Library Documentation center	**Library** UB **Public Library** VB **Documentation center**

Her har man, i en tesaurus, satt "Public Library" og "Documentation center" som synonymer for "Library." I den andre har man satt "Public Library" som underordnet term og "Documentation center" med en "se også"-henvisning. To forskjellige vurderinger av termenes forhold til hverandre er altså gjort. Begge er riktige, men de tilfører termene en litt forskjellig betydning. "At least we can conclude from this example that for any decision of conceptual exchangeability it is not sufficient to regard the entities as context-independent. The relationships within the structure of each knowledge representation must also be considered, since these relationships provide a major contribution to the meaning of the entities." (2014, s. 107-108).

I Thesaurus 1 i eksempelet er nærsynonymet "Documentation center" behandlet som et synonym. "Public Library" har egentlig en generisk relasjon til "Library," det er en del av begrepet og kunne vært plassert underordnet, som i Thesaurus 2. Likevel er det også behandlet som et synonym. "If near synonyms are treated as synonyms, the result set may contain ballast in the form of unprecise hits with respect to the search interest. The reason for this phenomenon is the merging of the conceptual differences of each near synonym for the purpose of representing the sum of meanings by one preferred term of the indexing language." (s. 238).

Denne "sum of meanings" er for eksempel vurderingen som er gjort i Thesaurus 1, hvor man har satt to termer som ikke egentlig er synonymer, som synonymer. Man har gjort en vurdering på at de er tilstrekkelig like den foretrukne termen til å sette dem som synonymer, fordi den foretrukne termen dekker det behovet man har for spesifisitet i samlingen tesaurusen benyttes til. Innbyrdes i Thesaurus 1 gir dette mening, og det gir mening til Library som foretrukken term. Men denne meningen er for det første ikke så lett å dokumentere, og heller ikke så lett å inkludere i mapping. Det vil også gi "Library" en litt forskjellig kontekstuell mening i Thesaurus 1 og Thesaurus 2.

Videre nevnes papegøyer som et eksempel. Papegøyer har flere plasseringer i Dewey. Det finnes både under biologi, og under husdyr. I SWD har termen "Papageien" ingen skiller mellom disse to betydningene. Hvis det benyttes "focused mapping", altså hvis man bare mapper til nummeret under biologi som betegner papegøyer generelt, vil man ikke kunne søke etter dokumenter indeksert med DDC som omhandler papegøyer som husdyr. Mapper man derimot med "comprehensive mapping" og tar hensyn til konteksten rundt Papageien i SWD, vil man kunne

mappe mot begge numrene (s. 109-111). En term fra kildevokabularet kan ha flere plasser i målvokabularet, som med papegøyer. Men det kan også være flere termer som passer til samme term i målvokabularet, ofte fordi termen er den mest spesifikke termen tilgjengelig, men fortsatt mer generell en termene i kildevokabularet. For eksempel inneholder SWD mange forskjellige typer ugler, som alle må eventuelt mappes til deweynummeret for ugler (s. 115-116).

Når det kommer til sammensatte emneord, både de som er naturlig sammensatt og de som blir satt sammen, altså prekoordineres, understrekes det at en forutsetning for interoperabilitet med disse er at både termene og relasjonene mellom dem må være være like for at de skal være identiske. For at man skal kunne avgjøre hvorvidt de sammensatte termene er identiske må også sammensetningen og meningen som ligger i denne være mulig å forstå."Any support by machines seems fictious as long as there is no transparency about the conceptual components." Å etablere interoperabilitet mellom sammensatte begreper vil være en krevende intellektuell oppgave som krever kjennskap til reglene for sammensetting i begge vokabuarene med mindre det som ligger i relasjonen mellom termene er tydelig (s. 134-136).

Når man mapper oppretter man relasjoner mellom emneord i forskjellige vokabularer. Disse relasjonene inneholder også mening. Noen relasjoner er allerede presentert i forbindelse med ISO-standarden. Disse relasjonene har sterk tilknytning til relasjonene vi tradisjonelt finner i en tesaurus. Men hva innebærer det for eksempel når man etablerer en relasjon som forteller oss at to termer er nesten like? Vi har allerede sett at, dersom man tar den kontekstuelle betydningen i betraktning, er svært få emneord helt identiske. Når man etablerer en relasjon som sier at "dette ligner, men ikke helt" får man ikke registert hva det er som gjør at de er nesten like.

En annen måte å uttrykke relasjoner på er tidligere nevnte "Degrees of deteminancy." Disse er sterkt preget av Dewey og mulighetene for nummerbygging. Et annet aspekt ved mappingrelasjoner er retningen. Dersom man etablerer at to termer er helt identiske, kan man benytte denne relasjonen begge veier. De fleste andre relasjoner går bare en vei (s. 137-139). Dette gjelder både relasjoner fra standarden og "Degrees of determinacy"

2.6 Andre prosjekter

De siste årene er det blitt utført mange forskjellige prosjekter knyttet til mapping av emneord. De fleste av disse er eldre enn standarden, slik at de er gjennomført med andre teknikker og typer mapping. Her presenteres et lite utvalg av prosjektene som finnes for å vise noen sentrale poenger ved mappingarbeid. Mappingen av Humord til WebDewey ansees også som spesielt sentral innen mapping i Norge i dag, men rapporten ble publisert for sent til å inngå i dokumentanalysen. Gjennom FinnOnto vises en litt annen tilnærming til interoperabilitet mellom emneordssystemer enn det ISO-standarden foreslår, og i det svenske emneordsarbeidet vises en annen måte mapping kan være nyttig på utover søk på tvers av systemer. Ytterligere fire prosjekter presenteres og diskuteres nærmere senere i oppgaven i forbindelse med analysen.

2.6.1 Humord mappet til Dewey

Som det ble nevnt innledningsvis pågår det i disse dager et arbeid med å opprette en generell norsk tesaurus. Et av delprosjektene utføres ved Universitetsbiblioteket i Oslo og kalles "Metodikk for mapping av Humord mot WebDewey." Prosjektet bygger på to tidligere prosjekter som presenteres nærmere i forbindelse med analysen, og kunne også vært høyst aktuell for analysen. Dessverre ble rapporten knyttet til prosjektet publisert for sent for å inkluderes der. Likevel er rapporten interessant, spesielt med tanke på formålet. Hovedmålet med delprosjektet er å "utvikle metodikk for datastøttet intellektuelt arbeid for mapping". (Gulbrandsen, Heggø, Knutsen & Seland, 2015, s. 3). Et ønske om å automatisere en størst mulig andel av mappingarbeidet er et forståelig ønske. Det bygger på en (rimelig) antagelse om at mye av mappingarbeidet kan gjøres av en datamaskin. "Vårt utgangspunkt har vært at datastøttet mapping vil danne et bedre utgangspunkt for å foreta en korrekt mapping enn dersom mappingen er en ren manuell prosedyre og dessuten også kreve mindre menneskelige ressurser." (s. 3).

I rapporten nevnes det, under overskriften "Hva er mapping og hvorfor er mapping nyttig?", at mapping gir muligheter for søk på tvers av emneordsvokabularer og klassifikasjonssystemer. Man kan benytte søketermer fra ett vokabular og få treff i dokumenter som er indeksert med et annet vokabular. Spesielt i en mapping til Dewey vil søkemulighetene være mange, fordi mange vokabularer på mange språk er mappet til nettopp Dewey i forskjellige utgaver (s. 4). Og selv om språkene er forskjellige, er jo numrene de samme.

I rapporten reflekteres det rundt en rekke potensielle utfordringer knyttet til arbeidet. Dette er som nevnt det tredje mappingprosjektet i rekken, så det er rimelig å anta at flere av de involverte har rikelig med erfaring knyttet til nettopp dette. Rapporten bemerker først og fremst at man mapper betydningsinnholdet i ordene. Selv om to emneord er like, har de ikke nødvendigvis det samme meningsinnholdet (s. 7). Dette er et sentralt poeng ved mapping. "I hvilken forstand kan vi si at Humord-emnet "arkitektur" og klassenummer 720 representerer det samme begrepet? Og hvis disse to representasjonene skal kobles sammen - hvordan skal da relasjonen mellom dem uttrykkes?" (s. 10).

Nærsynonymer og polysemi og håndteringen av disse nevnes som ett av fenomenene som kan føre til ulikheter mellom systemer. Det at termer eksisterer i en hierarkisk kontekst må også vurderes, ettersom dette også tilfører termen mening. Termer kan også være plassert forskjellig i forskjellige vokabularer. En av de grunnleggende utfordringene i nettopp mappingen mellom Humord og DDC er inndelingskriteriene for de forskjellige systemene. En tesaurus sorteres etter emne, mens et klassifikasjonsskjema sorteres etter fag. Denne forskjellen i struktur vil trolig føre til en-til-mange-mappinger i flere tilfeller.

Sorteringen gjør også at et fenomen nevnes ett sted i hierarkiet i Humord, mens det forekommer flere steder i Dewey. Her brukes eksempelet "Døden" som er plassert under helse i Humord. I Dewey forekommer "Døden" flere ganger, men tverrfaglige verker er plassert under sosiologi. I indekseringen er Døden i Humord benyttet innen flere fagområder. Et klassifikasjonsskjema har også mye informasjon knyttet til hvert nummer som er vanskelig å håndtere i et datasystem (s. 11).

En annen avgjørende forskjell på en tesaurus og et klassifikasjonssystem som Dewey er at en tesaurus baserer seg på postkoordinering, mens et klassifikasjonssystem baserer seg på prekoordinering. Dokumenter indeksert med Humord har i snitt fire forskjellige termer. Disse vil mappes til hvert sitt klassenummer. Rapporten foreslår at emneord i streng kunne vært lettere å mappe mot Dewey.

Det nevnes også utfordringer knyttet til menneskelige vurderinger som må gjøres og har blitt gjort. Å bedømme hvilke relasjonstyper som skal brukes er ikke alltid lett å avgjøre. Humord er også utstyrt med svært få topptermer, bare 26 stykker. Disse er forøvrig fagområder, et avvik fra vanlig praksis innen tesauruser. Likevel, på grunn av postkoordineringen, er termer bare plassert innen ett fagområde i Humord. Ser man videre på bruken av termen i indekseringen, kan den benyttes innen flere fagområder selv om den i tesaurusen bare har en plassering (s. 11-14).

Spesielt er mapping mellom tesaurus og klassifikasjonssystem preget av mange vanskelige avgjørelser fordi systemene er så ulike. Rapporten nevner at en mapping sjelden vil kunne uttrykkes som en eksakt likhet fordi klassebetegnelsene inneholder flere begreper. Dette til tross for at standarden hevder at denne likheten vil være den vanligste. Både det å velge hvilke termer som skal mappes og hvilken relasjon som skal opprettes mellom de to ansees som utfordrende. Det er nettopp derfor man ønsker å forbedre verktøy for automatisk mapping (s. 16-17).

2.6.2 FinnOnto

I Finland har en mengde aktører jobbet med et stort prosjekt knyttet til semantisk web siden 2003. Formålet med prosjektet, som har resultert i en mengde delprosjekter, er å etablere en infrastruktur for informasjon. “A solid, commonly shared infrastructure would make it much easier and cheaper for public organizations and companies to create interoperable, intelligent services on the coming semantic web. In our view, the infrastructure should be open source and its central components be maintained by the public sector in order to guarantee wide usage and interoperability across different application domains and user communities.” (Hyvönen et al., 2007, del 1).

Med et fokus på strukturering av emneordene som linked data skiller FinnOnto seg litt fra andre mappingprosjekter som ofte går ut på å få to vokabularer til å passe sammen. Blant annet gjennom å omstrukturere tesaurusen YSA til ontologien YSO ble emneordsvokabularet gjort semantisk meningsfullt for datamaskiner. For eksempel kan ikke en datamaskin forstå hvilken relasjon en “overordnet term”-relasjon betegner. Det gis følgende eksempel:
Halley’s comet BT Comet
Comet BT solar system

Halleys komet er en type komet. Komet er en av bestanddelene i et solsystem. Dette er to forskjellige typer forhold, som et menneske som ser på en tesaurus raskt oppfatter. For en datamaskin er det derimot umulig å forstå at Halleys komet er en bestemt komet og ikke en type kometer, eller at komet er noe som forekommer i et solsystem og ikke en bestemt type solsystem. Semantisk web har større muligheter for definering av relasjoner enn en tesaurus. En annen forskjell på RDF og en tesaurus er at det i RDF antas at en underordnet term også er underordnet alle termer oppover i hierarkiet. Altså at Halleys komet er en del av et solsystem. I dette tilfellet er det riktig, men i andre tilfeller stemmer det ikke. Et annet problem er at et emneord i en tesaurus kan inneholde flere betydninger, som et resultat av sortering på emne. Eksempelet som gis her er at emneordet Barn kan brukes om både en person som er barnet til noen og det kan brukes om en person som er i en tidlig fase av livet (del 3).

2.6.3 Svenska ämnesord og Dewey

I Sverige har man først i de senere årene begynt å benytte Dewey for klassifikasjon og hylleoppstilling. Tidligere var det et eget svensk system, SAB, som ble benyttet. Med valget om å gå over til Dewey kom også et behov for mappingarbeid. Kungliga biblioteks emneordssystem Svenska ämnesord (SAO) var tidligere knyttet til SAB, men nå måtte emneordene få deweynumre i stedet.

Overgangen ble lettet av det allerede eksisterende mappingarbeidet som var gjort. Det fantes for det første en tabell for overgang mellom SAB og DDC. Videre var SAO mappet til LCSH, som igjen er mappet til DDC. “Genom att mappningar mellan olika system redan finns bör möjligheterna till maskinellt stöd för detta arbete undersökas. Det finns mappningar från ca 70 % av termerna i Svenska ämnesord till LCSH. LCSH är i sin tur mappat mot DDC. På samma sätt bör också konverteringstabellen SAB-DDC kunna användas för maskinellt stöd vid mappningen Svenska ämnesord-DDC.” (Svanberg, 2006, s. 30-31).

Til forskjell fra de fleste andre mappingprosjekter hvor formålet er å gi emneordsvokabularet en slags tilleggsfunksjon i form av samsøk med andre samlinger eller lignende, var altså mapping et nyttig redskap i overgangen fra et klassifikasjonssystem til et annet i Sverige. I tIlegg til å basere

seg på og hente inn poster hvor det allerede fantes deweynummer, ble det vurdert å opprette et felles søk for poster med deweynummer og SAB-kode.

"Det tredje alternativet är att med hjälp av konkordansen SAB-DDC göra ett verktyg för samsökning av poster med SAB-kod och DDC-kod. Fördelen med denna metod är att alla katalogposter i LIBRIS kan göras sökbara. Nackdelen är att eftersom den bygger på mappningar mellan två klassifikationssystem med olika struktur och indelningar blir den mindre exakt. DDC-systemet bygger på att man väljer en kod per verk. Klassifikation med SAB innebär ofta att man använder flera klassifikationskoder. För att hantera dessa skillnader måste man hitta lösningar för hur poster med flera SAB-koder ska hanteras vid samsökning." (s. 37-39).

En slags postkoordinering av klassifikasjonskoder i SAB gir altså en utfordring man treffer på ved mapping av emneordssystemer hvor ett er basert på postkoordinering og ett på prekoordinering. Dette omtales nærmere i dokument- og dataanalysen.

3 Metode

3.1 Problemstilling

Selv om det finnes en standard for mapping av emneordsvokabularer, eller mer konkret en standard for mapping mellom en tesaurus og andre indekseringsspråk, er ikke mapping uproblematisk. Elementer som pre- og postkoordinerte emneord, hierarkiske strukturer og konteksten de brukes i og annen mening som tilføres emneordene er noen av faktorene som gjør at to like emneord ikke alltid er så identiske som man først skulle tro. Til tross for dette er mange mappingprosjekter både utført og under arbeid.

Som tidligere nevnt bygger denne oppgaven på en antagelse om at indekseringsspråk er ulike på en lik måte. Med utgangspunkt i dette er det ønskelig å forsøke å identifisere generelle utfordringer ved mapping gjennom å hente inn både tidligere erfaringer og å undersøke nye sett med emneord og deres mulighet for interoperabilitet for å undersøke fellesnevnere. Dette leder til følgende problemstilling:

Hva slags utfordringer kan oppstå ved forskjellige mappinger av indekseringsspråk?

- Hvilke fremgangsmåter er benyttet og hvilke utfordringer har oppstått i tidligere prosjekter?
- Hvilke fordeler og ulemper innebærer de forskjellige fremgangsmåtene?
- Hvilke fordeler og ulemper innebærer mapping av forskjellige typer indekseringsspråk?

Med "utfordring" menes her uønskede fenomener som oppstår i forbindelse med arbeidet. Utfordringene kan godt også kalles problemer, men utfordringene som omtales i denne oppgaven kan ikke nødvendigvis løses. I noen tilfeller antas det at utfordringene til en viss grad kan løses av egenskapene hos de forskjellige indekseringsspråkene, eller gjennom de forskjellige metodene. Disse eventuelle løsningene blir da ansett som fordeler, sammen med andre positive bidrag i mappingen som kommer fra fremgangsmåtene og indekseringsspråkene.

Som et motstykke kommer ulempene, som fører til utfordringer i mappingarbeidet. Det undersøkes mapping av indekseringsspråk, altså er både emneordsvokabularer og klassifikasjonssystemer inkludert. Hvorvidt noe er ønsket eller uønsket avhenger selvfølgelig av

hva man ønsker å oppnå. Derfor vies også formålene med prosjektene og mapping generelt oppmerksomhet.

3.2 Kvalitativ metode

Problemstillingen forsøkes besvart ved hjelp av en todelt metode. Analysens første del er en dokumentanalyse av publikasjoner tilknyttet fire forskjellige prosjekter hvor vokabularer med forskjellig struktur er blitt mappet ved hjelp av forskjellige teknikker og fremgangsmåter. Formålet med denne dokumentanalysen er å kartlegge erfaringer som er gjort i de forskjellige prosjektene, først og fremst hvilke typer utfordringer man har støtt på. Analysen vil basere seg på det som står skrevet i dokumentene.

Den andre delen av analysen er en dataanalyse hvor man forsøker å gjenskape og identifisere de samme utfordringene i en mapping av to nye vokabularer som ikke er blitt mappet tidligere, Humord og BIBBI. Her undersøkes det først og fremst hva en direkte mapping mellom vokabularene vil innebære. Der hvor det er relevant hentes også data fra Dewey inn.

Både dokumentanalysen og dataanalysen vil være kvalitative. Johannesssen, Tufte og Christoffersen skriver at "kvalitativ metode er særlig hensiktsmessig hvis vi skal undersøke fenomener som vi ikke kjenner særlig godt, og som det er forsket lite på, og når vi undersøker fenomener vi ønsker å forstå mer grundig." (2010, s. 32). Å gjennomføre en mapping av Humord og BIBBI uten noen form for analysering av resultatet ville vært lite fruktbart. Mapping i praksis er et relativt ferskt fenomen, og spesielt når de to utvalgte vokabularene er såpass forskjellige som de er, ville en kvantitativ undersøkelse hvor man forsøker å mappe større deler av vokabularer sagt oss noe om hvor mange mappinger som oppstår, men ikke hvorfor, og ikke noe om hva denne mappingen innebærer for informasjonen som allerede eksisterer og informasjonen som oppstår. Det ville heller ikke vært mulig å oppdage generelle utfordringer ved mapping gjennom et slikt arbeid.

Dataene fra dokumentanalysen vil være sekundærdata, altså allerede eksisterende data. Sekundærdata er blant annet egnet til å utforske andre menneskers erfaringer, meninger og praksiser (Braun & Clarke, 2013, s. 153). Som nevnt vil det være minst like gunstig nettopp å

undersøke andres erfaringer med mapping fremfor å gjøre dem selv. Innenfor det korte tidsrommet oppgaven skrives i, er det ikke like mye tid og ressurser til å sette seg inn i feltet som for eksempel en større prosjektgruppe har hatt tidligere. En sammenfatning av andres erfaringer er også en mangelvare innen fagområdet, og kan være til stor nytte. Både for fremtidige prosjekter med mapping, og for fremtidig forskning tilknyttet mapping.

Analysen vil forsøksvis kategorisere problemene og utfordringene ved mappingene. Det er uvisst om de samme typene problemer oppstår i mappingprosjekter med forskjellige utgangspunkt og metoder, men det er antatt at de vil det. Noen utfordringer er trolig unike for de forskjellige utgangspunktene, eller et større problem ved et utgangspunkt enn et annet. For eksempel er det mulig at store hierarkiske forskjeller er et større problem ved mapping mot et klassifikasjonsskjema.

3.3 Dokumentanalysen

Dokumentene som er valgt ut til dokumentanalysen er plukket ut fra 4 prosjekter som beskrives nærmere i neste kapittel. Det har blitt etterstrebet å finne prosjekter med forskjellig utgangspunkt og tilnærminger til mappingen, men med liknende formål for mappingen. To av de valgte prosjektene er norske og utført henholdsvis i 2013 og 2014. Dette betyr at de er utført etter at ISO-standarden er utgitt, og har dermed samme teoretiske utgangspunkt som dataanalysen i denne oppgaven. En mapping av terminologi på norsk har også vært avgjørende faktor i valg av prosjekter til dokumentanalysen. Prosjektene er også aktuelle da de danner et grunnlag for videre mapping av Humord til DDC, som er en del av arbeidet med en generell, norsk tesaurus. Interessant er også at de samme vokabularene, Realfagstermer og TEKORD, brukes i begge prosjektene.

Felles kildevokabular er også deler av motivasjonen for valget av MACS og CrissCross-prosjektet, som begge mapper tyske Schlagwortnormdatei (SWD), men i kombinasjon med vidt forskjellige vokabularer. I MACS får man undersøkt om mapping mellom flere språk har samme utfordringer som mappinger mellom samme språk, og i CrissCross mappes SWD til DDC, som igjen kan sammenlignes med mappingen av Realfagstermer til DDC.

Utvalget av dokumenter er begrenset til dokumenter på engelsk og norsk. Dette er trolig et tilstrekkelig stort utvalg, men det kan også hende at noe informasjon fra de tyske prosjektene som er valgt er bedre og mer detaljert presentert i tyskspråklige dokumenter. Dokumentene som er valgt er også sluttrapporter eller publisert etter at prosjektet er fullført. På en side gir dette uttalelsene i dokumentene en sikkerhet i form av at forfatteren har fått prosjektet litt "på plass" og ser helheten. På en annen side kan dette innebære at utfordringer som oppstod underveis, men ble løst på en enkel måte, ikke omtales, nettopp fordi det ikke lenger er nødvendig å diskutere eller ta forbehold om. På grunnlag av dette er det rimelig å anta at det med denne metoden ikke vil bli identifisert utfordringer som kan løses lett, og dermed heller ikke identifisere noen enkle løsninger på disse utfordringene.

Prosjektrapporter, artikler og liknende dokumenter er ikke typiske dokumenter som utsettes for dokumentanalyse. Likevel er en dokumentanalyse en nyttig metode for å skaffe oversikt over aktuelle utfordringer knyttet til mapping. Problemstillingen kunne også blitt besvart ved hjelp av intervju, men dokumentanalyse er valgt, blant annet på bakgrunn av en antagelse om at man har lettere for å formulere mer seg tydelig i et skriftlig format. Mappingarbeid kan ofte fremstå kaotisk og utfordrende, og det er først når prosjektene er fullført eller godt etablerte at man har oversikt nok til å peke ut virkelige, kanskje uløselige, utfordringer. I tillegg er mange utfordringer knyttet til mapping trolig langt lettere å forklare ved hjelp av eksempler og illustrasjoner fremfor løpende tekst eller tale.

3.4 Dataanalysen

Det er flere årsaker til at Humord og BIBBI ble valgt til dataanalysen. Den største og viktigste er en antagelse om at en slik mapping vil være svært utfordrende, om ikke umulig. Dermed er trolig tilfanget av utfordringer stort og godt illustrerende. Humord er en tesaurus, med henvisninger som hører en tesaurus til, og struktur som en tesaurus. Den er opprettet, vedlikeholdt og benyttet med formål om å indeksere en samling tilknyttet et stort fagbibliotek. Biblioteksentralen, på sin side, har en kundebase bestående nesten utelukkende av folkebibliotek, og oppretter, vedlikeholder og benytter emneordene sine med tanke på disse kundene - folkebibliotekene.

I tillegg er BIBBI strukturert som emneord i streng, og hvert emneord er tilknyttet et deweynummer. Slik kan man på den et vis si at BIBBI har en hierarkisk struktur i form av Deweys struktur, på et annet vis kan man kalle BIBBI et register til Dewey. I dataanalysen behandler man emnestrengens deweynummer som noe som tilfører strengene hierarkisk struktur. Som tidligere nevnt innebærer denne eksterne strukturen en sortering etter fag, ikke emne, som emneordsvokabularer tradisjonelt skal være sortert etter.

I dataanalysen er deweynumrene tilknyttet BIBBI hentet fra DDK5. På mange måter ville det vært langt mer nærliggende og aktuelt å benytte den norske oversettelsen av WebDewey. Flere av de omtalte prosjektene i dokumentanalysen er basert på mapping mot en oversettelse av WebDewey. Likevel er DDK5 valgt fordi det er her den helhetlige hierarkiske strukturen til BIBBI ligger. Bare deler av BIBBI er tilordnet WebDewey-numre i skrivende stund. Bruk av WebDewey ville dermed krevd en større arbeidsmengde da man måtte gjort antagelser om aktuelle deweynumre for både Humord og BIBBI-termer. Slike antagelser ville også gjort dataene noe mer upålitelige.

Poenget med dataanalysen er ikke å konstruere en realistisk mapping, men å undersøke hva som skjer når man forsøker å forene to så grunnleggende forskjellige systemer. I en videreføring av dokumentanalysen er formålet å undersøke og illustrere hvorvidt de identifiserte utfordringene er universelle for alle mappinger mellom forskjellige indekseringssystemer med forskjellig struktur, kontekst og innhold. Spesielt de grunnleggende forskjellene mellom prekoordinerte og postkoordinerte emneordssystemer og hvordan disse fungerer i forbindelse med mapping var også et interessant aspekt. Derfor er det blitt valgt ett prekoordinert og ett postkoordinert emneordssystem.

Dataene som behandles i analysen er hentet i perioden mars-juni 2015. Data fra Humord er hentet fra Humords nettbaserte søk. Data fra BIBBI er hentet via Biblioteksentralens interne katalogiseringsverktøy Promus.

Til tross for at dataanalysen ikke forsøker å være realistisk, vil den forhåpentligvis være nyttig. I arbeidet med en generell, norsk tesaurus foreligger Humord som grunnlag. Dersom man kunne

utnyttet noe informasjon fra BIBBI, et allerede eksisterende og veletablert vokabular som dessuten er generelt, kunne dette vært gunstig for tesaurusen. En eventuell mapping mellom en generell, norsk tesaurus og BIBBI kunne dessuten muliggjort søk på tvers av samlingene indeksert av Nasjonalbiblioteket og Biblioteksentralen. Valget av Humord og BIBBI til dataanalysen er ikke ment som en oppfordring om å mappe disse vokabularene direkte, da det er langt flere aspekter knyttet til et slikt arbeid enn de rent kunnskapsorganisatoriske som belyses i dataanalysen.

Dataanalysen er begrenset til innholdsbeskrivende emneord og -strenger, altså emneordene som plasseres i MARC-felt 650 hos Biblioteksentralen. Dette utelukker emneord plassert i andre felt, for eksempel personer, geografiske emner og formemner. Tilfellene som presenteres er valgt ut ved bevisst letning etter utfordringer. Flere tilnærminger for å identifisere utfordringer er benyttet. Først og fremst har det, der det er mulig og relevant, blitt gjenskapt konkrete tilfeller som er omtalt som eksempler i dokumentene fra dokumentanalysen. Noen av tilfellene er identifisert basert på egen erfaring fra arbeid med BIBBI som indekseringsverktøy, og spesielle utfordringer og særegenheter som er identifisert parallelt med arbeidet med oppgaven. Resten av tilfellene er identifisert ved ren utforsking og sammenligning av Humord og BIBBI. Tilfellene som presenteres i dataanalysen er dermed ikke dekkende for alle utfordringer som vil oppstå ved mapping mellom BIBBI og Humord, men eksempler på tilfeller som kan identifiseres med utgangspunkt i kategoriene fra dokumentanalysen.

3.5 Metodiske overveielser

Som tidligere nevnt falt valget på en dokumentanalyse blant annet fordi man i et skriftlig format får anledning til å formulere seg tydelig. Det antas også at man, innen man kommer så langt som å skrive en tekst som oppsummerer prosjektet, har en ganske god oversikt over prosjektenes styrker og svakheter. På en annen side er tekstene et ledd i å presentere prosjektene og resultatene for kollegaer og fagmiljø, og man er kanskje ikke så tilbøyelig til å diskutere alle svakheter i usminkede detaljer. Som tidligere nevnt er trolig problemer som ble løst underveis ikke nevnt, nettopp fordi de ble løst og dermed ikke trenger å presenteres.

En mer detaljert og usminket beskrivelse kunne muligens kommet frem gjennom intervju eller ved å få tilgang til prosjektdeltakernes notater og liknende dokumentasjon av prosjektets gang. Samtidig er ikke disse informasjonskildene like tydelige og godt oppsummerte som en tekst som skal publiseres. Analysen ville trolig også vært langt mer tidkrevende og dermed gått på bekostning av dataanalysen. Det er kanskje heller ikke så forskningsmessig interessant å undersøke problemer man allerede har løsningen på. Gjennom dokumentanalysen vil forhåpentligvis de mest sentrale og interessante utfordringene kunne avdekkes på en god måte.

Dataanalysen har også noen svakheter. For det første er det en svært liten andel av vokabularene som kan bli omtalt. Dermed er ikke dataanalysen egentlig noen god indikator på hva som vil skje ved en mapping av BIBBI og Humord. Til det ville man behøvd en langt mer kvantitativ tilnærming hvor man enten automatisk eller manuelt sammenlignet og forsøkte å mappe store deler av vokabularene. Et slikt arbeid kunne helt klart også vært interessant, men ville ikke egentlig svart på problemstillingen i denne oppgaven.

Det svært selektive datautvalget i analysen vil også føre til et særdeles dystert bilde av mapping. Dataanalysen er rett og slett problemorientert. Den er ment for å gjenskape og illustrere utfordringene som identifiseres i dokumentanalysen, og vil dermed ikke gi noen indikator på alle termene som potensielt kan mappes helt uproblematisk mellom de to systemene. Med utgangspunkt i den grunnleggende antagelsen om at man tenker nesten likt vil dette trolig være en stor andel. Dette får altså ikke dataanalysen påvist.

4 Om prosjektene

Selv om de fire prosjektene som er valgt er utført på forskjellige måter, med mappinger mellom svært forskjellige vokabularer, er målene med prosjektene like: å organisere dataene sine på en slik måte at man kan utveksle og samordne informasjon med andre. Dette er et mål som har blitt jobbet mot på mange måter og med mange typer data i mange år. Her følger en innføring av prosjektenes mål og metodene som er blitt benyttet for å forhåpentligvis kunne oppnå disse målene.

4.1 Realfagstermer + TEKORD

I et samarbeid mellom Realfagsbiblioteket ved Universitetet i Oslo og NTNU er det blitt gjennomført to prosjekter knyttet til mapping. Formålet med begge prosjektene var ikke mappingen i seg selv, men å undersøke metoder og belyse en rekke spørsmål knyttet til mapping. Det er likevel utført automatiske mappinger og intellektuelle vurderinger av kvaliteten på mappingene, slik at dataene fra de to rapportene disse prosjektene har resultert i, er interessante for sammenligning med andre mappingprosjekter. I det første prosjektet ble det undersøkt muligheter for mapping mellom de to bibliotekenes vokabularer. Dette resulterte i et lite, mappet vokabular kalt TORT. Det andre prosjektet bygget videre på det første, og undersøkte muligheten for mapping av de samme vokabularene til den norske oversettelsen av WebDewey. Dette omtales nærmere i del 4.2.

Formålet med det første prosjektet, som resulterte i rapporten "Realfagstermer og TEKORD" var "å undersøke om RDF [...] kan fungere som plattform for sammenlikning og sammenføyning av disse to emnesystemene som dekker realfag, naturvitenskap og teknologi." Deler av disse undersøkelsene bestod i å strukturere begge vokabularene i SKOS, for så å benytte et program som beregner tekstlig likhet mellom tekststrenger. Deretter ble det gjort en manuell gjennomgang av noen av ordene som ble oppgitt med en likhetsgrad mellom 90-100%. Gjennom denne informasjonen kunne man oppdage overlapp mellom de to vokabularene. Videre ble det skrevet en rekke spørringer for å avdekke informasjon om overlapp i foretrukne termer, synonymi og se også-henvisninger, for å avdekke om man kunne hente inn synonymer og henvisninger fra det andre vokabularet.

Det ble satt en rekke mål for prosjektet. Med et ønske om å berike begge vokabularene, ble det satt et mål om å avdekke overlappet mellom de to vokabularene. Det skulle også avdekkes om vokabularene kunne berikes med hverandres synonymer og se også-henvisninger, og om Realfagtermer kunne dra nytte av TEKORDS hierarkiske struktur. I tillegg ble det undersøkt hvorvidt TEKORD kunne dra nytte av Realfagstermers engelske termer. Et siste mål var å undersøke om det kunne hentes informasjon fra eksterne ressurser strukturert som linked data.

Formålet med å utvikle disse metodene var også å kunne bidra til kommende, liknende prosjekter. "På bakgrunn av det som er gjort internasjonalt, ønsker prosjektet å gi et bidrag til et langsiktig mål om å utvikle en samlet norskspråklig kontrollert emneordsbasert fagterminologi for store deler av kunnskapsuniverset i tråd med gjeldende språkpolitiske føringer." (s. 4). Som nevnt tidligere pågår det for tiden et prosjekt med formål om å mappe Humord til Dewey, som igjen er en del av arbeidet med en generell norsk tesaurus. Her er trolig undersøkelsene utført i dette prosjektet en nyttig erfaring. Tanken er at mapping skal være en del av informasjonsutveksling med både norske og internasjonale samarbeidspartnere.

"Ved å legge til rette for overganger mellom systemer ved hjelp av mulighetene som ligger i linked data og semantisk web vil deling og gjenbruk av data på tvers av systemer nasjonalt og internasjonalt kunne bli mulig, og informasjonssøkeren vil oppnå effektiv flerspråklig gjenfinning, emnebaserte navigeringsmuligheter og viderelenking til eksterne ressurser. Slik vil gjenfinningskvaliteten for både norske og fremmedspråklige informasjonssøkere ivaretas." (Kuldevere et al., 2013).

4.2 Felles terminologi for klassifikasjon med Dewey

I en videreføring av tidligere omtalte prosjekt undersøkte de samme institusjonene hvorvidt vokabularene Realfagstermer og TEKORD kunne benyttes til mapping mot Dewey. Hovedmålet for prosjektet var "å utrede mapping av terminologi i emneordsststemene Realfagstermer og TEKORD mot klasser i DDC". (Kuldevere et al., 2014, s. 4). Som i det foregående prosjektet er også dette en undersøkelse av hvorvidt en mapping er mulig, og hvilke metoder man kan benytte for å oppnå dette. Prosjektet skulle også undersøke muligheter for mapping mellom DDC og

UDC (Universal Decimal Classification), som TEKORD allerede er knyttet til. Det var også et mål å kvalitetssikre dataene fra foregående prosjekt.

I praksis ble Realfagstermer benyttet som kildevokabular, og emner som ble mappet til TEKORD (det som finnes i overgangen som kalles TORT) ble prioritert. Denne undersøkelsen vil blant annet benyttes til videre metodeutvikling i forbindelse med den planlagte mappingen av Humord til WebDewey. Som det forrige prosjektet var dette også et prosjekt hvor man utreder metoder for mapping, og ikke et faktisk mappingprosjekt. Likevel resulterte også undersøkelsene her i faktiske mappinger, og kvalitetssikring av disse. I dette prosjektet var formålet å finne metoder for automatisk mapping, og hvorvidt dette var en hjelp for intellektuell mapping. Det ble også forsøkt etablert fremgangsmåter for den intellektuelle mappingen.

Undersøkelsene ble utført gjennom en rekke tester for å finne riktig metode. Første test var en sammenligning av Realfagstermer med registertermene i Dewey. Denne testen avdekket utfordringer knyttet til konverteringen fra MARC21 til SKOS, og blant annet bygde numre og emneord i streng viste seg å være problematiske. Resultatene fra denne testen ble forkastet. I test to ble foretrukne og ikke foretrukne termer fra Realfagstermer sammenlignet mot klassebetegnelser og registertermer. Også her var strengene problematiske. Det ble ikke funnet noen treff ved forsøk på å benytte strenger, men de øvrige treffene var flere. Disse resultatene ble heller ikke jobbet videre med.

I test tre og fire ble foretrukne og ikke-foretrukne termer fra Realfagstermer mappet mot klassebetegnelsene i DDC, og ikke registertermene. Metoden var den samme for begge testene, men test fire inneholdt oppdaterte data fra Realfagstermer og inkluderte 600-gruppa fra DDC, som inneholder anvendt vitenskap som medisin og ingeniørfag og dermed må kunne kalles sentral for mapping fra et vokabular beregnet for Realfag. Økningen av data førte også til økning av antall mappinger, og totalt 1681 mappingforslag fra test 4 ble behandlet videre. Disse automatisk genererte mappingene ble igjen vurdert videre ved hjelp av en systematisk sjekk opp mot blant annet informasjonen knyttet til klassebetegnelsen i Dewey og også ved å se på bruken av emneordet i katalogen (Kuldevere et al., 2014).

4.3 MACS

I et samarbeid mellom Library of Congress (USA), Deutche National Bibliothek (Tyskland), Bibliothèque nationale de France, British Library og det sveitsiske nasjonalbiblioteket ble det igangsatt en mapping mellom engelske, franske og tyske emneord. De engelske emneordene ble hentet fra Library of Congress Subject Headings (LCSH), de franske fra Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU), og de tyske fra Schlagwortnormdatei (SWD). Formålet med prosjektet var "to create a multilingual gateway (not a multilingual thesaurus) that allows subject access to library catalogues in order to overcome linguistic barriers." (Jahns & Karg, 2010, s. 51).

Selv om det å koble opp de respektive institusjonenes samlinger mot hverandre var et mål, som i de andre prosjektene, skiller MACS seg noe fra disse ved at motivasjonen stammer fra språklige forskjeller. Formålet er at et emneord på ett språk skal fungere som en inngang til samlinger indeksert på tre forskjellige språk. Til forskjell fra de fleste mappingprosjekter er det heller ikke valgt noe kildevokabular eller målvokabular, alle vokabularene ble vurdert som likestilte. Dette innebærer også at det er opprettet mappinger som går flere veier mellom vokabularene, slik at tre forskjellige begreper kan bli registrert som helt like. "The mapped concepts, represented by the different subject headings, have to be truly close equivalents to provide successful multilingual subject retrieval. The quality of linking is based on the retrieval of consistent sets of bibliographic records from the different catalogues." (s. 58).

Det foreslås at mappingene blant annet kan brukes til å hente inn informasjon fra katalogposter registrert i de involverte landene. Også brukerne kan dra nytte av mappingene, ved at de kan søke i flere samlinger uten å måtte kjenne språket alle dokumentene er indeksert med. Flerspråklige innganger kan også være en støtte for brukere av bibliotek som ikke har språket biblioteket benytter som morsmål.

4.4 Criss Cross

CrissCross-prosjektet er et todelt mappingprosjekt med hvor man i begge prosjektene mapper med SWD, men mot forskjellige vokabularer. Det er det tyske nasjonalbiblioteket som står bak. En del av Criss Cross-prosjektet bestod i mappingen av SWD til LCSH og RAMEU i MACS, som nettopp har blitt omtalt. Den andre delen av prosjektet bestod i mapping av SWD til den

tyske versjonen av Dewey. For enkelhets skyld omtales mappingen av SWD til Dewey heretter som CrissCross, mens den andre delen av prosjektet omtales som MACS, sammen med de andre mappingene som ble utført i forbindelse med MACS. Det er altså en mapping av SWD til Dewey som omtales her.

Formålet med CrissCross-prosjektet var " to create a thesaurus-based and user-friendly research vocabulary to facilitate search in heterogeneously indexed collections by linking topical headings of the German subcject headings authority file Schlagwortnormdatei (SWD) to notations of the Dewey Decimal Classification (DDC)." Med andre ord vil en mapping til DDC gi en mulighet for å søke i flere samlinger. Videre skal CrissCross forbedre søkeprosessen og sørge for kontinuerlig bruk av dataene som allerede finnes. Mappingene omtales som verktøy for å forbedre søkeprosessen, både for å gi bedre tilgang til DDC i seg selv, for å kunne rangere søkelister og for å kunne gi felles innganger til forskjellig indekserte samlinger.(Jacobs, Mengel & Müller, 2010, s. 37).

Som tidligere nevnt opererer CrissCross med andre relasjonstyper enn standarden (se kap. 2.4). Disse "Degrees of Determinacy" er en av tre retningslinjer som karakteriserer prosjektet. De to andre er en "en-til-mange-strategi" som går ut på at man mapper et emneord fra SWD til flere klassenumre i DDC. Dette er en god løsning på en utfordring knyttet til forskjellene mellom emneordssystemer og klassifikasjonsskjemaer, og illustreres og diskuteres videre i de kommende delene. Den tredje retningslinjen er "Deep level mapping", som ligner litt på spesifisitetsprinsippet. Sammensatte begreper ble alltid mappet til det mest spesifikke nummeret, og numre ble bygget dersom det var nødvendig og tillatt (s. 38-39).

4.5 Oppsummering

Til tross for noe ulike tilnærminger og vidt ulike vokabularer som utgangspunkt, jobber alle prosjektene mot svært lignende mål. Alle har et ønske om å kunne strukturere dataene sine på en slik måte at de kan være nyttige for andre institusjoner, samtidig som man kan dra nytte av andre institusjoners data. Ord som "samordne" og "utveksle" står sentralt. Denne samordningen skal både være fordelaktig for indekserere, som kan hente inn andres indekseringsarbeid, og for brukere, som kan få mulighet til å søke etter dokumenter via emneordsinnganger på tvers av

samlinger og indeksseringsspråk. Dette gjelder både innefor en nasjons språk og i et flerspråklighetsperspektiv. Spesielt mapping via Dewey blir gjort med et mål om at det vil knytte en institusjons emneord opp mot et internasjonalt miljø hvor flere mappinger blir gjort.

Selv om det på detaljnivå er store ulikheter mellom prosjektene, er det altså bare snakk om forskjellige fremgangsmåter for å oppnå samme mål. Målet er i dette tilfellet først og fremst en mulighet for å søke i flere samlinger indeksert med forskjellige emneordsvokabularer. Utfordringene som diskuteres i analysen er derfor i stor grad knyttet til ulikheter i vokabularene som på en eller annen måte forringer gjenfinningseffektiviteten.

Gjennom å undersøke hvorvidt automatisk mapping kan være en hjelp for intellektuell mapping i de norske prosjektene finnes det også et mål om å spare ressurser. Selv om en fullstendig intellektuell mapping vil gi færre feil, krever det store ressurser. Dessuten er det også i intellektuell mapping en rekke utfordringer. Disse drøftes nærmere i dokumentanalysen.

5 Dokumentanalyse

Dokumentanalysen ble utført på totalt fem dokumenter. Ved gjennomlesing ble det lett bevisst etter setninger og avsnitt som kunne knyttes til mål og formål for prosjektene, utfordringer og ulikheter knyttet til mappearbeidet, og eventuelle løsninger eller fordeler ved disse ulikhetene. Ikke alle ulikheter ved mapping er nødvendigvis ulemper eller blir omtalt som utfordringer. De er likevel inkludert i analysen. Etter gjennomlesing ble mål, utfordringer og fordeler fordelt i kategorier.

Det ble identifisert sju hovedkategorier av problemer som var til stede i større eller mindre grad ved de forskjellige metodene. De kan oppsummeres slik:

- Homonymi
- Sammensatte begreper
- Forskjellige hierarkiske plasseringer eller grader av inndeling
- Forskjellige begreper som et resultat av forskjellig bruk og kontekst ved indeksering
- Støy
- Konsistens i forbindelse med menneskelig vurdering av relasjonene
- Ressursbruk

Av disse kategoriene er det to typer utfordringer knyttet til selve arbeidet. Dette er konsistens ved kontroll av vurdering av relasjonene og ressursbruk. Videre er det fem kategorier som kan knyttes til oppbygning og sortering av de forskjellige systemene.

Fire av disse, homonymi, hierarki, sammensatte begreper og utfordringer knyttet til forskjellig bruk og kontekst, er knyttet til sortering og struktur i de allerede eksisterende systemene og kan trolig undersøkes nærmere i dataanalysen. Støy ved automatisk indeksering er også et resultat av sortering, men kan i likhet med de to resterende kategoriene, konsistens ved kontroll av mappinger og ressursbruk, reduseres gjennom arbeid med de fire kategoriene som undersøkes nærmere i dataanalysen.

I kategorien for homonymi havnet ganske enkelt homonymer. Ord som staves likt eller nesten likt kan identifiseres som like ved mappingforslag som genereres basert på ordlikhet, men ved

nærmere ettersyn er de to emneordene betegnelser for vidt forskjellige begreper. I kategorien for sammensatte begreper ble det plassert begreper som ikke kan betegnes med ett enkeltstående emneord, eller som ikke har blitt betegnet med ett enkeltstående emneord i ett av vokabularene. Utfordringer knyttet til strenger ble også plassert her. Innunder hierarki-kategorien ble utfordringer og forskjeller knyttet til hvordan emneordene plasserer seg hierarkisk i forhold til hverandre plassert. Spesielt i forbindelse med mapping av et emneordssystem til Dewey, som er et klassifikasjonssystem, oppstod det utfordringer knyttet til hierarki.

Utfordringer knyttet til kontekst og forskjellig bruk er en kategori som inneholder noen utfordringer som også kan knyttes til hierarki og sammensatte begreper, men fellesnevneren her er at konteksten emneordene er oppstått og benyttet i gjennom mange år har tilegnet emneordene noe forskjellig mening som kan føre til feil ved mapping, til tross for at emneordene fjernet fra denne konteksten godt kunne blitt mappet. Her er også utfordringer knyttet til intern praksis og bruk av henvisninger plassert. Denne informasjonen om emneordene er heller ikke dokumenterbar ettersom den generelt sett skjer over tid og mer eller mindre ubevisst.

I kategorien støy, som er en relativt liten kategori, ble utfordringer knyttet til rene feilkoblinger plassert. Denne kategorien inntraff først og fremst ved automatisk mapping. Disse utfordringene diskuteres lite, og resultater med mye støy forkastes som regel. Som tidligere nevnt kan trolig støy reduseres noe ved arbeid på de fire kategoriene nevnt over. I tillegg kan bedre verktøy for konvertering av data og lignende trolig hjelpe, men dette arbeidet faller utenfor denne oppgavens tematikk.

Utfordringer knyttet til konsistens ved vurdering av mappinger var lette å identifisere, da flere forfattere påpekte rett ut og konkret at dette arbeidet var svært vanskelig å gjøre på en effektiv og nøyaktig måte. Ressursbruk var også en type utfordring som ble omtalt, om enn på en mer udefinerbar måte. Forskjellige aspekter ved arbeidet med mapping, eller arbeidet i sin helhet, ble omtalt som krevende, tidkrevende eller utfordrende, og videre arbeid med data ble i flere tilfeller valgt bort på grunnlag av at det var krevende.

En utfordring i forhold til ressursbruk er vanskelig å konkretisere da "for mye arbeid" avhenger av hvilke forventninger man har om mengden arbeid som vil oppstå, samt hvor mye tid og øvrige ressurser man har til rådighet i prosjektet. Likevel er det trolig et universelt ønske om å utføre mappingarbeidet på en effektiv måte med få feilkilder, slik at man oppnår best mulige resultater på kortest mulig tid.

5.1 Homonymi

Homonymi var en utfordring ved prosjektene som inneholdt en automatisk generering av relasjoner. Ord med stor likhet ble koblet sammen, men ved nærmere ettersyn og kontroll opp mot ting som emneordets hierarkiske plassering, bruk og synonymer, viste det seg at det dreide seg om homonymer.

I mappingen av Realfagstermer mot DDC (Kuldevere et al., 2014) oppstod det flere utfordringer knyttet til homonymi. I noen tilfeller er ordets betydninger enkle å skille, som når belgfrukten linser i DDC foreslås mappet med Linser i Realfagstermer, som omtaler linser i kameraer og lignende. Her har man både deweynummer og en hierarkisk plassering i Dewey som sier noe om emneordets betydning. Andre er vanskeligere, som Antenner, som ikke hadde noen informasjon knyttet til seg i Realfagstermer. Her måtte man se på hvordan begrepet var brukt for å avgjøre hvorvidt det var snakk om teknologi eller biologi.

Som det nevnes i mappingen av Realfagstermer og TEKORD (Kuldevere et al., 2013), er utfordringene med homonymi knyttet til automatikken. I dette prosjektet ble alle ord med 100% likhet akseptert som like. Her ligger det trolig endel homonymer, men det er rimelig å anta at andelen er noe mindre enn hvis den samme metoden hadde blitt benyttet på to generelle vokabularer. Ettersom Realfagtermer og TEKORD dekker nærliggende fagområder er det rimelig å anta at ord som ser like ut også beskriver det samme begrepet, snarere enn det begrep fra et helt annet fagområde. En lignende tankegang tas videre i mappingen av Realfagstermer og DDC, hvor det diskuteres enkelte fordeler ved at det bare ble mappet mot deler av DDC.

"Når hele Dewey-oversettelsen er ferdig, og vi får tilgang til den, vil vi få mappingforslag mot klasser i andre grupper enn de vi har hatt tilgang til så langt. Noen av RT-begrepene som nå er registrert i verktøyet vårt med kun én kobling mot DDC, vil få flere. [...] Kanskje vil det vise seg

at avgrensning til noen hovedgrupper i Dewey-materialet faktisk er det mest fornuftige for å unngå for mye støy, men vi vil isåfall måtte gjøre en grundigere vurdering av hvilke grupper som bør inngå i analysen." (Kuldevere et al., 2014, s. 20). Videre undersøkelser med begrensning av fasetter som mappes er helt klart interessant. Samtidig kan dette også begrense antallet forskjellige faglige vinklinker som tilknyttes et begrep.

Ren homonymi, som i eksempelet med Linser, kan løses ved bruk av kvalifikatorer. Det nevnes blant annet i forbindelse med mappingen av Realfagstermer og TEKORD som et av forbedringspotensialene at "I tillegg vil bruk av kvalifikatorer på homonymer være svært nyttig." (Kuldevere et al., 2013, s. 16). I et "lite" mappingforsøk som dette, som likevel inneholder tusenvis av begreper, er det kanskje mulig å knytte kvalifikatorer til problematiske emneord, men dette er nok en manuell jobb som eventuelt må gjøres. En slik praksis forutsetter også en gjensidig innsats og konsistens på potensielle homonymer fra begge institusjonene.

I MACS oppstod det også utfordringer knyttet til polysemi, hvor det samme ordet betyr *nesten* det samme, som i eksempelet med Cidre/Cider/Apfelwein. En amerikansk Cider og en fransk Cidre er ikke det samme, selv om begge deler er drikke laget hovedsaklig av eple. I akkurat dette tilfellet ble det vurdert at i et indekseringsforemål var termene like nok til å kunne registreres som like (Jahns & Karg, 2010, s. 13). For noen brukere vil søk via denne mappingen føre til ny informasjon om andre typer epledrikke, mens det for andre vil gi støy i form av typer epledrikke de ikke er interessert i. Denne formen for utfordringer er også knyttet til bruk og kontekst, som omtales senere.

5.2 Sammensatte begreper

Alle prosjektene hadde utfordringer knyttet til sammensatte begreper, men de forskjellige vokabularene fikk også litt forskjellige utfordringer med forskjellige løsninger. I MACS gav forskjellige regler for orddeling utfordringer da tysk har flere sammensatte ord enn engelsk og fransk. Utfordringen ble løst ved bruk av boolske operatorer, nærmere bestemt operatoren AND. Dette ligner i stor grad på sammensatt ekvivalensmapping som foreslås i ISO-standarden. Et eksempel som gis er det tyske begrepet "Leichtathletiktrainer." Et tilsvarende begrep fantes i LCSH; "Track and field coaches". Fra RAMEAU ble det konstruert et nytt begrep ved hjelp av AND: "Athlétisme AND Entraîneurs" Denne løsningen er i midlertidig ikke uproblematisk, da

MACS i utgangpunktet baseres på at alle tre vokabularer mappes likt, og at (nesten) alle relasjoner er ekvivalensrelasjoner, som dermed kan gå begge veier. Bruken av AND vanskeliggjør denne toveis-bruken av relasjonene (s. 59). Det er mulig at lignende utfordringer kan oppstå ved mapping av norske emneordsvokabularer til for eksempel engelske, da norsk språk også har større muligheter for sammensatte ord, dog ikke like store som det tyske språket.

I mappingen av Realfagstermer og TEKORD inneholdt datamaterialet blant annet 632 strenger (Kuldevere et al., 2013, s. 6). Ettersom metoden bestod i å sammenligne begreper med en viss tekstlikhet, er det tilsynelatende ikke gjort noen mapping av sammensatte begreper som er formulert forskjellig. Det er derimot etterlyst kvalifikatorer knyttet til enkelte begreper for å unngå homonymi.

Også i mappingen av Realfagstermer til Dewey ble det utelatt å undersøke muligheter for mapping av sammensatte begreper. Det diskuteres i midlertidig hvorvidt Dewey er godt egnet som målvokabular ved mapping av strenger. "En mapping mellom RTs emnestreng Utdødde arter : Fugler og DDCs registerterm Utdødde fugler kunne for eksempel vært opprettet basert på at alle ordene i den korteste strengen finnes i den lengre strengen. Kanskje kan det også være gunstig å kutte ned ordene til ordstammene sine («stemming») før sammenlikning." (Kuldevere et al., 2014, s. 19).

I testrunde 1 og 2 oppstod det utfordringer. I test 1 til bygde deweynumre som gav "tilsynelatende uforklarlige treff" og i test 2 ble det forsøkt å inkludere strenger i søk, men de gav ingen treff (s. 8). Det kan virke som om måling av tekstlikhet alene ikke håndterer bygde numre og strenger så godt. Det foreslås en metode hvor man splitter strengene. Dette er den samme som foreslås i standarden, men ved å splitte strengene forsvinner også informasjonen som ligger i at nettopp disse ordene er kombinert i en bestemt rekkefølge.

I mappinger til Dewey er det ofte begrep i Dewey som er "for sammensatt", slik at de lengste bygde numrene umulig kan ha en match i et vokabular. Spesielt med vokabularer basert på postkoordinering vil ordene passe inn på et mer generelt nivå. "Whereas DDC concepts are mainly pre-combined and discipline-focused, many SWD headings do not exhibit a specific (a

priori) context. Such a context is only constituted a posteriori in the act of subject indexing when several SWD headings are combined to represent a document's topic, following post-coordination rules." (Jacobs, Mengel & Müller, 2010, s. 25).

Selv om de sammensatte begrepene ikke ble mappet fra Realfagstermer til Dewey, bød Deweys prinsipper for bygde numre på utfordringer. Det var blant annet et problem å mappe begrepet Terapi i Realfagstermer, fordi det i Dewey føyes til Terapi til forskjellige lidelser. Ved postkoordinering kombineres den enkeltstående termen Terapi med den aktuelle lidelsen, mens det i Dewey gjøres gjennom nummerbygging. Det bemerkes at begrepet Terapi i Realfagstermer ansees som et emneord som bør benyttes i kombinasjon med andre emneord (Kuldevere et al., 2014, s. 12). Med andre ord er formålet det samme, men uttrykksmetodene blir forskjellige.

Utfordringene knyttet til Deweys nummerbygging er også å spore i hierarkisk plassering. Dette diskuteres nærmere i neste avsnitt. Ved gjenfinning vil effekten av dette være at man ikke får hentet fram dokumenter med lange deweynumre ved søk på emneord fra postkoordinerte vokabularer. Disse er mappet til kortere, mer generelle numre, selv om emneordene i bestemte kombinasjoner kanskje kunne passet på mer spesifikke dokumenter. Ved prekoordinering vil trolig denne utfordringen forekomme sjeldnere.

5.3 Hierarki

Utfordringer knyttet til hierarki oppstod i alle prosjektene utenom MACS hvor alle mappingene er til identiske begreper. Slike mappinger er reversible og den hierakiske strukturen spiller ingen rolle ved mappingen.

I mappingen av Realfagstermer og TEKORD var ikke utfordringene så store, trolig fordi Realfagstermer ikke har en hierarkisk struktur. Det ble derimot funnet at Realfagstermer kunne hente en god del struktur fra TEKORD. Det bemerkes at også hierarkiske relasjoner må vurderes manuelt. Det er ikke utredet hvordan strukturen eventuelt skal tas inn i vokabularet (Kuldevere et al., 2013, s. 17). Det er heller ikke drøftet hvordan gjenfinningen eventuelt påvirkes. Skulle en term få tilordnet flere nye underordnede termer, bør disse underordnede termene velges dersom spesifisitetsprinsippet følges.

En konsekvens av dette kan være at noen av dokumentene som omhandler temaet tilknyttet den underordnede termen får det mest spesifikke emneordet, mens dokumenter som ble indeksert før de underordnede termene ble innført vil være tilknyttet den overordnede termen. Det samme kan oppstå ved mapping av to vokabularer hvor ett går mer i detalj på et område, mens det andre har klart seg med den overordnede termen.

Ettersom deweynumre i mange tilfeller dekker et større fagområde enn enkeltstående ord, vil ofte mappinger til Dewey resultere i noe støy ved søk. Flere ord mappes til ett nummer - og dermed vil søk på dette nummeret resultere i treff både knyttet til begrepet man leter etter, men også til andre begreper som igjen er knyttet til nummeret. I Criss-Cross foreslås det at denne effekten kan begrenses noe ved å rangere søk etter hvilke typer likhet som er registrert - hvor stor grad av likhet (Jacobs, Mengel & Muller, 2010, s. 42-43).

Det er mange utfordringer knyttet til mapping til Dewey og hierarki. Ett begrep kan ha flere plasseringer i hierarkiet i Dewey, alt etter vinkling. I tillegg innebærer de fleste numre mer enn ett begrep, fordi det i Deweys opprinnelige formål, hylleoppstilling, er to begreper som er tilstrekkelig like til at de kan stå sammen på hylla. I praksis vil de fleste termer i for eksempel en tesaurus få flere mappinger til forskjellige numre, og hvert nummer vil få flere termer fra tesaurusen. I CrissCross-prosjektet fant de at mange av termene fra SWD var mer spesifikke enn Deweynumrene som passet best, slik av mange termer ble mappet til et Deweynummer med lavere spesifitet. Det påpekes at denne effekten bidrar til å samordne ellers svært smale begreper. Både i trefflister og i navigering av SWD kunne dette være til hjelp for sluttbrukeren (Hubrich, 2010, s. 77-79).

I mappingen av Realfagstermer mot Dewey ble det som tidligere nevnt bare mappet mot deler av Dewey. Selv om det også her oppstod en del tilfeller av at ett begrep hadde flere mappinger mot Dewey, bemerkes det at det trolig blir enda flere i en eventuell mapping mot hele Dewey. De mange plasseringene som er tilgjengelig i Dewey bød også på utfordringer tilknyttet nærsynonymer, som for eksempel da begrepet Knokler ble foreslått mappet til både knokler hos dyr og knokler hos mennesker. I Realfagstermer innebærer begreper begge typer knokler, mens i

Dewey er det plassert to steder. Begge relasjonene ble valgt betegnet som tilnærmet eksakt ekvivalens (Kuldevere et al., 2014).

5.4 Kontekst og forskjellig bruk

De fleste mappingprosjekter startes opp med et ønske eller mål om å forenkle samordning og utveksling av informasjon med andre institusjoner. Behovet for forenkling stammer fra det faktum at forskjellige indekserere arbeider under forskjellige forutsetninger med forskjellige formål og brukergrupper som benytter seg av arbeidet de gjør. Denne grunnleggende ulikheten i institusjonenes arbeid, og dermed også i emneordsvokabularene som oppstår i takt med at samlingen vokser og indekseringsbehovene endrer seg, gjør at hvert vokabular med tiden kan få noen særegenheter og "unntak fra reglene." Disse gjør seg godt og er svært fordelaktige i hver enkelt samling, men byr på utfordringer når man forsøker å sette likhetstegn mellom eget vokabular og andres.

Disse utfordringene finner vi blant annet i eksempelet hvor nærsynonymer i TEKORD var oppgitt med BRUK-henvisninger til en foretrukket term. Synonymer i seg selv er først og fremst en berikelse for systemene ved mapping, men nærsynonymer som i hvert system er behandlet som synonymer, kan by på feilaktige relasjoner.

Noen "Brukt for"-relasjoner i TEKORD viste seg å ikke være direkte synonymer, men snarere assosiative relasjoner som kanskje heller burde vært representert med "Se også"-henvisninger. I indeksering innen en samling er det, for eksempel i dette tilfellet ved bruk av TEKORD, gunstig å kunne gi indekserer og bruker beskjed om at "vi bruker ikke Demninger, se under Dammer", men når Demninger er foretrukket term i Realfagstermer, vil Dammer feilaktig bli identifisert som synonym for Demninger (Kuldevere et al., 2013, s. 11-12).

Gjennom indekseringen kan emneord også få en betydning knyttet til seg. En naturlig del av indekseringsprosessen for de fleste vil være å undersøke hvilke dokumenter man allerede har knyttet til et emneord, nettopp for å i størst mulig grad samordne dokumenter med samme innhold. Dette er emneords formål i praksis. Men gjennom en slik arbeidsmetode vil det hele tiden blir gjort vurdering på hva et emneord innebærer, og disse vurderingene vil være forskjellige fra indekserer til indekserer, og fra emneordsvokabular til emneordsvokabular. I

noen tilfeller vil en slik forskjell i betydning gi en høyere grad av fullstendighet ved søk. Man får hentet inn flere dokumenter med tilnærmet lik betydning, som av en eller annen grunn ikke er vurdert som lik nok av forskjellige indekserer i en gitt tid og en gitt situasjon. Samtidig kan det også kombinere emneord som med tiden er vurdert noe ulikt i en så stor grad at det har gitt emneordene en litt annen betydning.

Et enkelt eksempel på dette er det tidligere nevnte problemet med begrepet Antenner i mappingen av Realfagstermer til DDC. Det var ingen annen informasjon tilknyttet begrepet Antenner i Realfagstermer. Kun gjennom å se på hvordan emneordet hadde blitt benyttet i katalogen var det mulig å avgjøre at det var nettopp snakk om antenner designet for å motta signaler, og ikke antenner på insekters hoder. Hvorvidt begrepet var utelukkende tiltenkt denne betydningen da det ble innført som emneord, vites ikke. Tidligere nevnt er også begreper knyttet til epledrikk, omtalt med emneordene Cider, Cidre og Apfelwein. Her er det kulturelle forskjeller som gir emneordene litt forskjellig betydning. Kulturelle forskjeller trenger ikke nødvendigvis oppstå over landegrenser, to vitenskapsretninger kan like mye ha en forskjellig vinkling eller tolkning av det samme fenomenet.

I det tidligere nevnte eksempelet med Knokler var emneordet tolket som knokler hos både mennesker og dyr i Realfagstermer, mens det i Dewey var tolket som to forskjellige ting. Løsningen her var å mappe emneordene til tilnærmet like. I en eventuell direkte mapping av for eksempel et vokabular benyttet til en veterinærmedisinsk samling og et benyttet til en samling rettet mot medisinsk personell ville trolig emneordet Knokler i hvert av vokabularene være tilknyttet såpass forskjellige dokumenter at en slik mapping ville ført til støy.

Ved en mapping mellom to vokabularer vil sannsynligvis konsekvensene av forskjellig bruk av lignende begreper være mer omfattende. Dewey er, i motsetning til de fleste emneordssystemer, ikke utviklet og i konstant utvikling i tilknytning til alle dokumentene den benyttes til. Mens emneordsvokabularer har informasjon tilknyttet dem i form av bruk, har Dewey en felles tilleggsinformasjon uavhengig av bruk i form av noter, forklaringer, inkluderer-noter, anvisninger for bruk, og så videre.

I mappingen av Realfagstermer til Dewey blir det påpekt at det bør utvises forsiktighet ved mapping av emneord av allmenn karakter, for eksempel Metoder, Utvikling og Vekst. Ettersom disse emneordene ofte benyttes i kombinasjon med andre emneord, som vist i det tidligere nevnte eksempelet med Terapi, er det problematisk å mappe dem til et bestemt fagområde i Dewey. "Bare ved å sjekke faktisk bruk kan vi finne ut om slike emneord er brukt i spesiell kontekst, eller om de sprer seg over flere fagområder. RT har ikke noe strukturelt skille mellom innholdsbeskrivende emneord [...] og innholdsbeskrivende emneord av allmenn karakter [...]. Dette er en mangel ved vokabularet som kan forklares ved hvordan det er laget, og som det vil ta tid å rydde opp i." (Kuldevere et al., 2014, s. 12).

Rapportene som er gjennomgått i denne dokumentanalysen fokuserer først og fremst på utfordringene knyttet til selve mappearbeidet mellom to vokabularer. Selv om gjenfinning og søk trolig alltid ligger i underbevisstheten når valg tas og utfordringer kommenteres, er ikke selve søkene hvor man anvender de mappede systemene omtalt. Slike undersøkelser vil kanskje avdekke flere utfordringer knyttet til forskjellig bruk. Med mindre man etter mappingen går sammen om et felles emneordsvokabular og en felles indekseringspraksis, er det også en risiko for at emneordene etter at mappingen er utført får forskjellig mening gjennom forskjellig bruk med tiden. Dette er det tatt høyde for i CrissCross-prosjektet, hvor hver mapping får et "time stamp" som indikerer når mappingen ble opprettet (Hubrich, 2010, s. 80). Dette garanterer ikke for en konsistent bruk av emneordene, men gjør det enklere å oppsøke utdaterte mappinger.

5.5 Støy

Støy oppgis som en utfordring i de norske prosjektene. De automatiske metodene som blir benyttet gir av og til resultater som er feil uten at det er noen tydelig forklaring på hvorfor den feilaktige mappingen oppstod, eller hvordan man kan løse det. Disse mappingene må ganske enkelt forkastes.

Jo mer data man inkluderer i mappingen - jo flere feil, og jo mer å luke ut. Det er trolig på grunn av dette at prekoordinerte emneord og sammensatte begreper i så stor grad er utelatt fra prosjektene. "Generelt vil utvidelser av metoden innføre mer støy i form av ikke-relevante forslag, så en vil hele tiden måtte vurdere mengden relevante forslag mot mengden støy."

(Kuldevere et al., 2014, s. 19). Spesielt med tanke på at Realfagstermer bare ble forsøkt mappet mot 500-gruppen og 600-640 i Dewey, er det muligheter for både flere treff og mer støy ved senere mapping mot hele Dewey.

I samme rapport understrekes det at man bør undersøke metoder for å redusere støy. Det foreslås å gjøre en frekvensanalyse, og å unngå innholdsbeskrivende emneord av allmenn karakter. Å sammenligne foreslåtte mappinger basert på ordlikhet mot bruk av emneord og klassifikasjonsnummer i katalogen kan godt mulig være en god metode for å begrense støy. Samtidig kan en slik sammenligning muligens også luke ut noen relevante mappinger som bare ikke forekommer så ofte, og dermed minke presisjonen noe.

5.6 Vurdering av relasjoner

På samme måte som alle de tidligere nevnte utfordringene har oppstått som et resultat av at mennesker tenker og vurderer ulikt, fant man også flere utfordringer knyttet til det å vurdere og å etablere relasjoner. Både i prosjektene hvor hver relasjon ble opprettet som et resultat av intellektuelle vurderinger og der hvor relasjoner først ble opprettet automatisk og deretter ble vurdert intellektuelt er det umulig å vurdere helt likt. "Alt tyder på at mapping og vurdering av relasjoner er en subjektiv prosess når den skal utføres manuelt/intellektuelt. Det kommer vel neppe som noen overraskelse når vi vet hvor ulikt man kan sette emneord og klassifikasjon på ett og samme dokument. Desto viktigere blir det å utarbeide retningslinjer som kan minimere subjektiviteten." (Kuldevere et al., 2014, s. 21).

Som de tidligere omtalte utfordringene viser, blant annet med tanke på homonymi, er en feilfri, helautomatisk mapping foreløpig ikke en mulighet. Det må derfor tas høyde for at det alltid vil være en utfordring knyttet til menneskelige vurderinger, og at det trolig er umulig å mappe to vokabularer på en "perfekt" måte som alle kan være enige om.

Utfordringer her kan minskes noe ved å etablere rutiner for kontroll og å sørge for at mer enn én person vurderer hver relasjon. Dette vil igjen være ressurskrevende. I mappingen av Realfagstermer mot Dewey sjekket "kontrollørene" først informasjon knyttet til vokabularene, deretter sjekket de bruk i katalogen, og så ble type relasjon valgt. Hvis to kontrollører velger samme relasjon, er den godkjent.

5.7 Ressursbruk

Selv om prosjektene har svært forskjellig grad av manuell vurdering av relasjoner, er det i alle prosjekter bemerket at disse vurderingene er ressurskrevende. Hvorvidt mengden ressurser som kreves er et problem avhenger til en viss grad av hvilke ressurser, og ikke minst midler, man har tilgjengelig, men det er klart at et effektivt arbeid alltid er ønskelig. I mappingen mellom Realfagstermer og DDC stilles det spørsmål ved hele prosjektet:

"Er det i det hele tatt mulig å mappe to så ulike systemer? Man kan velge å innta et enkelt standpunkt - nei, det er ikke mulig. Prosjekt slutt. Eller man kan være pragmatisk og prøve. Vil man prøve, er det nye spørsmål å svare på: Hva er hensikten? Hvor skal nytteverdien finnes: I indekseringsfasen, i søkefasen - eller i begge ender? Er det et poeng å angi relasjoner for flest mulig emneord, uansett hvor snevre treff de har mot klassenumrene - eller kommer det til et nivå der vi ser at en sammenheng finnes, men vi anser at å etablere den vil skape støy? " (Kuldevere et al., 2014, s. 20).

Det er trolig verdt å vurdere disse spørsmålene før man i det hele tatt setter i gang. Mappingen er tilsynelatende et krevende arbeid, både i form av å få emneord som bør mappes til å mappes, og i å gjøre vurderinger i forhold til emneord som bare kanskje burde mappes. Spesielt i tvilstilfeller kan en bevissthet rundt formålet med arbeidet være avgjørende for å begrense ressursbruk. Det er selvfølgelig mulig å for eksempel velge å bare opprette mappinger for fullstendig ekvivalens for å redusere antallet tvilstilfeller og mengden arbeid, men dette vil igjen gi mindre utveksling av informasjon.

Til tross for at det bare ble benyttet mappinger for fullstendig ekvivalens i MACS, var det også her et svært omfattende arbeid. Alle mappinger ble vurdert manuelt, på tvers av tre språk. Riktig nok ville ikke en sammenligning av tekstlikhet virket ved mapping på tvers av forskjellige språk, slik at en manuell mapping var nødvendig i akkurat MACS-prosjektet. "Compared to cross-concordances of other projects, we can agree that it is a cost-intensive and time-consuming effort to generate such terminology networks. But we are convinced that it is a good investment and

that the data created can be used as a training database for automatic systems." (Jahns & Karg, 2010, s. 61).

I mappingen mellom Realfagstermer og Dewey gis det også uttrykk for at vurderingsarbeidet er krevende å gjøre på en konsistent måte, som nevnt tidligere. Likevel viser de andre utfordringene som er omtalt her at slike vurderinger er nødvendige for å oppnå et godt resultat med tilstrekkelig god gjenfinningseffektivitet. Selv når man bare mapper ord som er helt identiske, vil det oppstå utfordringer knyttet til homonymi, hierarkisk plassering og forskjellig bruk.

Med utgangpunkt i dette er det en rimelig påstand at mapping, inntil videre, er et ressurskrevende arbeid. Å oppnå en viss grad av automatisk genererte relasjoner vil i de fleste tilfeller være ønskelig og vil kunne lette arbeidet. Løsninger til utfordringene knyttet til automatisering vil dermed også kunne lette arbeidsmengden noe.

5.8 Oppsummering

Det ble identifisert sju kategorier av utfordringer i dokumentene. Fem av dem er knyttet til strukturelle og språklige utfordringer ved selve vokabularet, og to er knyttet til selve mappingarbeidet. Utfordringene er til en viss grad knyttet til hverandre. Forskjellig bruk av vokabularene og konteksten emneordene oppstår i gir strukturelle forskjeller. Disse forskjellene gir utfordringer i tilknytning til homonymer, sammensatte begreper og vokabularer. Utfordringer knyttet til ressursbruk, konsistens ved valg av relasjoner, og støy kan minskes ved å løse eller minske utfordringene knyttet til strukturelle forskjeller mellom vokabularene og forskjellig bruk.

Utfordringer som oppstår i flere kategorier samtidig vil trolig ofte oppstå ettersom et emneord får tilført mening fra flere forskjellige kilder. Det oppstår og benyttes alltid i en kontekst, det har som regel en hierarkisk plassering i vokabularet det opprettes i, og det har, hvis det er nødvendig, kvalifikatorer knyttet til seg ved tilfeller av homonymi eller består av flere ord dersom det er et sammensatt begrep. På tross av disse "mulitkategoriske" utfordringene var det mulig i dokumentanalysen å betegne de fleste identifiserte utfordringer med en bestemt kategori.

Som forventet var forfatterne av de analyserte dokumentene noe tilbakeholdne med å vise fram feil og mangler ved arbeidet sitt. Det var likevel tilstrekkelig omtale av utfordringer som oppstod

i de forskjellige prosjektene, og disse hadde en såpass stor likhet at de var mulig å fordele i de sju kategoriene. Det ble ikke tolket frem noe av tekstene utover det som sto skrevet. Ved nærmere undersøkelse av de tilgjengelige resultatene og dataene etter prosjektene kunne det trolig blitt identifisert flere utfordringer, men rapportene dekket i stor grad de samme utfordringene slik at det er rimelig å anta at det meste av utfordringer er nevnt i større eller mindre grad.

Ikke alle ulikheter mellom emneordsvokabularene ble omtalt som ufordelaktige. For eksempel ble nummerbygging og dermed høyere spesifisitet i Dewey trukket fram som en fordel i CrissCross-prosjektet, med eksempelet "Jakt" fra SWD, som da kunne inndeles finere i flere typer jakt ved mapping mot Dewey. I motsatt tilfelle, hvor CrissCross er mer spesifikt enn Dewey, kan det brukes til å samle smale begreper (Hubrich, 2010, s. 77-78).

Selv om utfordringene har noen fellesnevnere er det tilsynelatende både fordeler og ulemper ved de forskjellige metodene, og forskjellige typer vokabularer bestemmer også størrelsegraden av utfordringene. Å mappe alt gjennom intellektuelt arbeid er naturlig nok ressurskrevende, men løser noen utfordringer knyttet til blant annet homonymi. Å gjøre alt automatisk krever færre ressurser, men gir utfordringer i tilknytning til blant annet støy og homonymi.

En potensiell utfordring som ikke er omtalt er synonymi. Det finnes en mulighet for at samme begrep er betegnet med to forskjellige ord, og dermed ikke identifiseres. Det er også viet lite oppmerksomhet til emneord i strenger i rapportene, selv om flere av vokabularene som er mappet inneholder prekoordinerte emneord. Det diskuteres likevel hvorvidt emneord i streng kunne vært enklere å mappe mot Dewey. Bruk av kvalifikatorer foreslås som et tiltak for å begrense utfordringer knyttet til homonymi. Det er også en mangel på undersøkelser av og diskusjon om hvordan forskjellige vokabularer forholder seg til hverandre ved mapping til et felles målvokabular. Utfordringene identifisert knyttet til forskjellig bruk og emneordenes kontekst kan tyde på at dette er verdt videre undersøkelser.

6 Dataanalyse

I dataanalysen vil noen av de identifiserte kategoriene fra dokumentanalysen eksemplifiseres med emneord fra Humord og BIBBI emneord. Der det er relevant omtales deweynumre fra DDK5 som allerede er knyttet til emneordene eller –strengene fra BIBBI. Vurderingene som er gjort er først og fremst med tanke på hvilken effekt utfordringene vil ha ved et eventuelt søk på tvers av samlinger. Det er dette formålet som oppgis oftest og står sterkest i de omtalte mappingprosjektene i oppgaven. Samtidig vil utfordringene ha en negativ innvirkning for de fleste formål med mapping, slik at det uansett er relevante problemstillinger.

Dataanalysen får ikke belyst alle utfordringene som ble identifisert i dokumentanalysen. Ressursbruk og konsistens i valg av relasjoner er vanskelig å måle, da dataanalysen er et svært begrenset arbeid utført av en person. Alle sammenligningene gjøres intellektuelt, og det er dermed utfordrende å identifisere tilfeller hvor en får automatisk foreslåtte mappinger som kan kategoriseres som støy. Å gjøre en intellektuell identifisering av og vurdering av mappinger i dataanalysen vil også utelukke noen utfordringer som står sterkere i en automatisert mappingprosess. Som vi har sett i dokumentanalysen er de fleste hovedkategoriene av utfordringer problemer som krever en intellektuell vurdering. Dermed er denne tilnærmingen også mest nærliggende i dataanalysen. Selv uten forsøk på automatisering av mappingene er det mulig å identifisere rikelig med forskjeller i de ulike vokabularene som gir anledning til å vise noen utfordringer knyttet til hierarki, sammensatte begreper og forskjellig bruk, samt diskutere utfordringer knyttet til homonymi.

6.1 Homonymi

I mappingen av Realfagstermer og TEKORD ble det foreslått å benytte kvalifikatorer for å begrense tilfellene tilknyttet homonymi. Nettopp i forbindelse med homonymi er prekoordinerte emneord som BIBBI gunstige, fordi fagområdet kan oppgis. I kombinasjon med deweynummer er det få muligheter for forveksling. Skulle man derimot, slik standarden foreslår, splitte strengene til enkeltstående emneord og så forsøke å mappe dette direkte til Humord, vil det trolig oppstå endel tilfeller av homonymi. Det vil også oppstå endel homonymer innad i BIBBI, fordi konteksten de er tilknyttet fjernes. Ved slike utfordringer er likevel de opprinnelige strengene nyttige for å enten utelukke mappinger hvor emneordet har mer enn ett ledd, hvor ledd nummer

to ikke matcher eller ved å kontrollere disse tilfellene. For eksempel er homonymet "boksere" skilt med en tilføyelse i parentes, og med forskjellig deweynummer i BIBBI: "Boksere (Hunder): 636.73" og "Boksere (Sportsfolk): 796:83092"

Det har ikke vært mulig å identifisere noen tilfeller av homonymi mellom BIBBI og Humord gjennom metoden som er valgt for dataanalysen. Dette kan tyde på at prekoordinerte emner gir færre tilfeller av homonymi. Skulle man velge å bare mappe det første emneordet i hver streng, ville det derimot oppstått en rekke tilfeller. Det kan også være tilfelle at en automatisk generering av mappinger basert på ordlikhet mellom BIBBI og Humord ville avdekket enkelte homonymer. Men gjennom manuell gjennomgang var det ikke mulig å identifisere noen tilfeller av homonymi.

6.2 Sammensatte begreper

Ettersom alle emnestrenger i BIBBI får et deweynummer i forbindelse med at de opprettes, er lengden og formen på strengene noe preget av dette. Der hvor det er rom for å bygge lange numre blir det noen ganger også rom for lange strenger, som vanskelig kan samordnes med enkeltstående emneord. For eksempel strenger som denne:

Landssvik - Norge - Kunstnere (historie) : 948.1053

Som tidligere nevnt ligger det mye informasjon mellom linjene i slike emnestrenger. Som mennesker forstår vi at Landssvik - Norge - Kunstnere (historie) betyr "historiske dokumenter om landssvikende kunstnere i Norge." Det ville vi også forstått om de fire ordene i strengen var enkeltstående emneord. Ved en mapping fra BIBBI til Humord er det lite trolig at man ville kunnet mappe med en likhetsrelasjon. Derimot kunne det vært en mulighet, gjennom å splitte strengene som standarden foreslår, å mappe alle strenger som begynner med "landssvik" som smalere emner under Landssvik i Humord. Humord har også termen Landssvikere, som trolig vil beskrive flere av de samme dokumentene som strengen. Her må det gjøres en intellektuell vurdering om man skal mappe til en av termene i Humord, eller begge.

I mappingen av Realfagstermer til Dewey oppstod det en utfordring i forbindelse med emneordet Terapi, ettersom dette heller tilsluttes forskjellige sykdommer enn å være et enkeltstående

begrep. Det ble også bemerket i rapporten at man anså Terapi som et emneord som bare burde benyttes i kombinasjon med andre emneord. En lignende utfordring er tilfellet i Humord og BIBBI, ettersom BIBBI henter struktur fra DDK5 og dermed sorteres etter fag framfor emne. Både BIBBI og Humord inneholder emneordet Terapi, men mens det i Humord benyttes i kombinasjon med de forskjellige tilstandene som skal behandles i hvert dokument, knyttes den aktuelle tilstanden som behandles og Terapi sammen i strenger i BIBBI.

Ved en kikk på hvilke dokumenter de to identiske emneordene er knyttet til, blir forskjellen på prekoordinerte og postkoordinerte emneord svært tydelige. Ved en mapping fra BIBBI til Humord ville derfor alle de 125 dokumentene om forskjellige typer terapi indeksert med Humord bli koblet sammen med disse 5 dokumentene:

ISBN	Forfatter	Tittel	År
87-7728-040-7	Ryce-Menuhin, Joel	Jungiansk sandleg ; den virkningsfulde terapi / Joel Ryce-Menuhin ; oversat af Jytte ...	cop.1993
82-7413-346-3	Höting, Hans	Qigong-kuler ; magiske helsekuler fra Kina gjør deg frisk og glad! / Hans Höting ; over...	cop. 1995
82-05-09342-3	Staff, Peer Hallvard	Trening av pasienter / av Peer H.Staff, Sverre Mæhlum og Kåre Rodahl	1978
91-87680-26-2	Hu Bin	Övningar i traditionell kinesisk terapi / Hu Bin ;översättning Niclas Bengtsson	1993
		The Quiet roar [video] / director Henrik Hellström ; scriptwriters Henrik Hellström, Fredr...	2015

Uten å anta for mye om qigong-kuler og jungiansk sandlek er det neppe det de fleste får anbefalt av fastlegen. Ettersom de fleste former for terapi passer inn i Dewey på et mer spesifikt nivå, er det kun oversiktsverker og terapiformer som er så spesielle at de ikke passer inn andre steder som har blitt indeksert med selve emneordet Terapi i BIBBI.

6.3 Hierarki

Et klassisk eksempel på en term som befinner seg innen flere fagområder er Stein. Også i BIBBI er Stein plassert mange steder, med kvalifikator bak.

Stein : arkitektur : 721.044
Stein : billedhogging : 736.5
Stein : formingshobbyer : 745.58
Stein : landskapsarkitektur : 717
Stein - Dekorasjonsmaling : 745.723
Stein - Norge : byggematerialer : 691.209481

Alle disse emnene handler om Stein som materiale i forskjellige sammenhenger. Generelle verker om forskjellige typer stein vil heller finnes under emneordet Bergarter med deweynummer 552. I Humord er Stein underordnet Geologi.

Stein

Brukt for:

Steinsorter
Overordnet term:
Geologi
Toppterm:
Realfag
Se også:
Naturstein (Materiale)
Her følger underordnete termer på alle nivåer:
-Edelsteiner
--Diamanter

I tillegg er det en se også-henvisning til Naturstein som materiale. Ved en mapping til Humord vil trolig den beste løsningen være en mange-til-en-mapping hvor alle strengene som begynner med Stein mappes til emneordet Stein i Humord. Ettersom dette emneordet trolig er kombinert med de forskjellige fagområdene ved indeksering, vil sannsynligvis Stein i Humord innebære mye av de samme som alle stein-strengene i BIBBI kombinert. Ulempen her er at presisjonen knyttet til søk mot de enkelte fagområdene indeksert i BIBBI senkes noe, mens fullstendigheten trolig vil øke.

I dokumentanalysen ble det diskutert hvorvidt også det at mange emneord hører sammen innunder samme deweynummer kan by på utfordringer. BIBBI inneholder emneordene "Feministisk økonomi" og "Økonomi og feminisme", begge knyttet til deweynummer 330.082. Nummeret er bygget, av nummeret for Samfunnsøkonomi og hjelpetabell-nummeret for kvinner, men som det også ble nevnt i dokumentanalysen er det nettopp nummerbyggingen i Dewey som kan gjøre den mer egnet for mapping av sammensatte begreper. Det er en forskjell på de to BIBBI-emneordene. Feministisk økonomi er et eget forskningsfelt innen økonomi, og emneordet omfatter da dokumenter om dette feltet. Økonomi og feminisme kan brukes til å beskrive dokumenter som omhandler disse to fenomenene i en sammenheng. Trolig vil publikasjoner innen feministisk økonomi ofte også passe med emneordet Økonomi og feminisme. Om man

skal følge spesifisitetsprinsippet, bør man velge Feministisk økonomi. Likevel har de to emneordene såpass lik betydning at de har et felles deweynummer.

Humord har både emneordet Feministisk økonomi, samt emneordene Økonomi og Feminisme. Dermed blir mapping av Feministisk økonomi en smal sak, men hva med BIBBIs Økonomi og feminisme? Man kan splitte det opp og mappe til henholdsvis Økonomi og Feminisme i Humord, og håpe at disse to termene brukt sammen dekker mye av de samme dokumentene som "Økonomi og feminisme" i BIBBI. Men vil en postkoordinering av Økonomi og Feminisme uttrykke det samme som emneordet Økonomi og feminisme?

Dette eksempelet er også knyttet til sammensatte begreper. Det er blitt tatt en avgjørelse hos Biblioteksentralen om å etablere emnestrengen "Økonomi og feminisme" for å uttrykke nettopp bøker som omtaler økonomi i et feministisk perspektiv. Dette er noe annet enn hvis man etablerte emnestrengen "Feminisme og økonomi" som da uttrykker at feminisme er viktigst, og trolig ville omhandlet feminisme i et økonomisk perspektiv. Denne forskjellen fremkommer ikke ved bruk av postkoordinerte emneord. Trolig er begge vinklinger relevant for de fleste brukere, men man mister altså muligheten til å skille mellom de to. Samtidig får man gjennom postkoordineringen samlet dokumenter med begge vinklingene. Men sidestillingen av Feministisk økonomi og Økonomi og feminisme i BIBBI vil vanskelig kunne ivaretas ved en mapping til Humord, til tross for at de to begrepene er svært nærliggende hverandre.

I dette tilfellet er det bare snakk om to emneord tilknyttet samme deweynummer. I mange tilfeller kan det være langt flere emneord som bør mappes mot samme nummer, og dermed flere komplikasjoner. Men det kan også være omvendt - at ett emneord har flere deweynumre. For eksempel er Samer plassert mange steder i BIBBI avhengig av hvilken vinkling dokumentene har:

Emne
Samer : 948.00494
Samer : 305.89455
Samer : eiendomsrett til landområder - økonomi : 333.2
Samer : menneskerettigheter : 323.119455
Samer : minoritetssosiologi : 305.89455
Samer : norsk statsforfatningsrett : 342.087
Samer - Forskning : minoritetssosiologi : 305.89455
Samer - Fortellinger : 305.89455 &
Samer - Fotografisamlinger : 779.2
Samer - Humor : minoritetssosiologi : 305.8945507
Samer - Canada : 971.0049455
Samer - Canada - Historie : 971.0049455
Samer - Evenes - Kulturhistorie : 948.44430049455

Til og med Samer som enkeltstående emneord har to forskjellige deweynumre. Her ser man også veldig godt konsekvensene av å sortere etter fag fremfor emne. Noen av emnene havner under samfunnsfaglige aspekter i 300-gruppen, noen havner innunder geografiske aspekter og fotografisamlinger samles under kunst. Her er det formen som får bestemme den hierarkiske plasseringen og ikke emnet. I mappingen av Realfagstermer mot Dewey ble det bare mappet mot deler av vokabularet, og det ble diskutert at en slik begrensning kunne være fordelaktig for å begrense perifere mappinger. I tilfeller som Samer, som i BIBBI er plassert i tre av Deweys ti hovedgrupper, kan et slikt valg fort redusere fullstendigheten.

Videre viser de to enkeltstående emneordene "Samer" i BIBBI godt hvordan et emneord innebærer mer enn selve ordet. Her er ordene tilført en meningsforskjell i form av numrene. Bare noen med tilgang til et eksemplar av DDK5 eller med inngående kjennskap til systemet vil være i stand til å forstå forskjellen på de to emneordene. I søk er det uproblematisk å kombinere de to emneordene, men i indekseringsprosessen vil det være en forskjell. I Humord har man Samer som et enkeltstående emneord, underordnet Folkegrupper og overordnet noen forskjellige grupper med samer. I tillegg henvises det til Samepolitikk:

Samer
Brukt for:
Kystsamer
Brukt for:
Lapper (Folk)
Brukt for:
Sjøsamer
Brukt for:
Fjellsamer
Brukt for:
Samefolket
Brukt for:
Skogsamer
Brukt for:
Sami
Overordnet term:
Folkegrupper i Europa #
Toppterm:
Folkegrupper
Se også:
Samepolitikk
Her følger underordnete termer på alle nivåer:
-Kolasamer
-Lulesamer
-Sørsamer
-Østsamer
--Skoltesamer

Også her ser man hvordan enkelte nyanser kan gå tapt. Man kan velge å mappe begge betydningene av Samer i BIBBI til Samer i Humord, enten som smalere begrep eller som en tilnærmet likhet. Men forskjellen som ligger de to deweynumrene blir vanskelig å bevare. På samme måte som med eksempelet Landssvik kan man velge å mappe inn alle strengene som begynner med ”samer” som smalere begrep under Samer i Humord.

6.4 Kontekst og forskjellig bruk

I dokumentene var det flere utfordringer knyttet til måten vokabularene og emneordene har blitt benyttet og strukturert på. Forskjellig praksis er etablert etter behovet til samlingen og brukerne innenfor institusjonene som har vokabularene. Ofte ser man bort fra reglene for vokabularet man har for å forbedre gjenfinningen innen egen samling. Ulike praksiser fører fort til ulikheter mellom vokabularene, som lett kan føre til feilkoblinger og forringet gjenfinning. Et godt eksempel på dette innen Humord og BIBBI er begrep og emneord knyttet til demens. Mens

medisinske bibliotek, som trolig har størst behov for emneord knyttet til demens, benytter seg av den medisinske tesaurusen MeSH (Medical Subject Headings), er dokumenter om demens også en del av innholdet i samlingene som dekkes av Humord og BIBBI. Her er det et behov for et mer generelt detaljnivå, og behovet er dekket gjennom løsninger som er svært forskjellige.

I BIBBI er emneordene knyttet til demens, som alle andre emneord i BIBBI, organisert etter Dewey:

Demens = 616.83
Demens: sosialmedisin = 362.19683

Alzheimer = 616.83
Alzheimer : sosialmedisin = 362.19683

Senil demens : geriatri = 618.976898
Senil demens : geriatri : sosialmedisin = 362.198976898

Her har selve sykdommen blitt plassert sammen med andre sykdommer, i 616.83. Dette nummeret deles blant annet med emneordet Alzheimer, som er en av de vanligste årsakene til demens. På samme måte er demens og alzheimer innen sosialmedisin plassert på samme nummer som er bygget. Senil demens er betegnelsen på demens hos pasienter som er over 65 år, og dette er plassert innunder geriatrien. I Humord er det valgt en annen, mer samlingsspesifikk løsning.

Senil demens

Brukt for:
Aldersdemens
Brukt for:
Senil dementia
Brukt for:
Alderdomssløvsinn
Brukt for:
Alzheimers sykdom
Brukt for:
Pick-Alzheimers sykdom
Brukt for:
Senilitet
Brukt for:
Demens
Overordnet term:
Psykiske lidelser
Toppterm:
Helse

Her er Senil demens foretrukket term for blant annet Demens, mens Senil demens i realiteten er en type demens. I tillegg er det også en Brukt for-henvisning til Alzheimers, som i BIBBI er likestilt hierarkisk med Demens. Ved en direkte mapping fra BIBBI til Humord ville dermed alle distinksjonene som er gjort i BIBBI mellom de forskjellige typene Demens, samt Alzheimer, forsvunnet. Enten velger man å ikke mappe begrepene, og dermed faller informasjonen ut, eller man kan velge å betegne Demens som tilsvarende Senil demens i Humord. Ved en mapping faller forskjellen man har gjort mellom Demens og Senil demens i Bibbi ut, da begge emneord mappes til Senil Demens i Humord.

6.5 Oppsummering

Selv om dataanalysen er basert på intellektuelle vurderinger, er eksemplene som nå er blitt presentert basert på ord og termer med høy bokstavlikhet. Dermed kunne de trolig også blitt foreslått som automatiske mappinger. Felles for eksemplene er at de krever en vurdering av emneordenes kontekst og mening for å kunne velge en best mulig løsning for hvert eksempel. Som det ble funnet i dokumentanalysen er slike vurderinger en av utfordringene ved mappingarbeid, fordi det er krevende å gjøre på en konsistent måte. Selv om eksemplene her er valgt for å teste ut kategoriene fra dokumentanalysen, er det mye som tyder på et en eventuell mapping fra BIBBI mot Humord ville krevd en stor innsats innen intellektuelle vurderinger. Dette ville igjen ført til en ny kategori av utfordringer, i form av stor ressursbruk.

I flere av eksemplene ble det foreslått å mappe BIBBI-strengene mot en mer overordnet term i Humord. Mye tyder på at det vil være lettere å opprette en viss fullstendighet ved å mappe kompliserte, sammensatte begreper mot en overordnet, enkeltstående term enn det vil være å opprettholde presisjon ved mapping fra prekoordinerte emner mot en tesaurus. Det er mulig at en mapping av begge vokabularene mot Dewey som et navn vil være bedre egnet for å opprettholde en viss presisjon. Dette er midlertidig ikke blitt undersøkt i denne dataanlysen, men vil være interessant videre forskning når en mapping fra Humord til WebDewey, og eventuelt BIBBI til WebDewey, foreligger.

7 Diskusjon

Det er fort gjort å bli nedslått av funnene som blir presentert i dokument- og dataanalysen. Ved å lete etter utfordringer og problemer, har resultatene dannet et noe negativt preget bilde av mapping som noe usedvanlig utfordrende. Slik er det heldigvis ikke i et faktisk mappingarbeide. De aller fleste termer vil sannsynligvis relativt uproblematisk kunne tilordnes en av mappingtypene som omtales i ISO-standarden. De ulikhetene som vises i analysen, er ikke nødvendigvis alltid utelukkende negative.

Innledningsvis ble det gjort en antagelse om at det vil oppstå generelle utfordringer ved mapping av emneord. Mulige utfordringer har blitt presentert og drøftet gjennom tidligere forskning, dokument- og dataanalyse. I diskusjonen sammenstilles disse funnene og diskuteres ytterligere, i et forsøk på å svare på problemstillingen og forskningsspørsmålene. De identifiserte kategorienes tilstrekkelighet diskuteres også, da de gjennom dataanalysen viste seg å være noe utfordrende å bruke. En avgjørende faktor for hvilke utfordringer som oppstår viste seg å være strukturen i vokabularene man velger å mappe, og hvilke relasjoner man velger å mappe med. Disse utgangspunktene diskuteres også.

7.1 Finnes det generelle kategorier av utfordringer?

I analysen ble det identifisert sju kategorier av utfordringer. Fire av disse ble bragt videre i dataanalysen, og anvendt her med eksempler som illustrerer at kategoriene kan antas å være av en allmenn karakter, og dermed mulige å anvende for å forutse potensielle utfordringer ved de fleste mappingarbeider. Dette kan være forskjellige feilkoblinger og mappinger som vil redusere presisjon eller fullstendighet, eller forårsake støy. Det må nevnes at BIBBI og Humord, vokabularene som ble valgt for dataanalysen, ble valgt blant annet på grunnlag av at de er svært ulike i struktur og formål, og ville vært svært arbeidskrevende å mappe i et faktisk mappingprosjekt Sånn sett kan dataanalysen fremstå som litt virkelighetsfjern. Likevel tjente den svært godt formålet med å undersøke hvorvidt kategoriene av utfordringer identifisert i dokumentanalysen var av en allmenn karakter.

Tre av kategoriene som ble identifisert i dokumentanalysen var ikke mulig å illustrere direkte i dataanalysen. Likevel antas det at forbedringer innen utfordringene knyttet til indekseringsspråks

system og sortering, altså de fire kategoriene som ble benyttet i dataanalysen, vil redusere omfanget av de tre andre. Med færre problemer kommer også færre vanskelige tilfeller av mapping som må vurderes manuelt, det vil forhåpentligvis redusere ressursbruken og kan også redusere automatisk genererte mappinger som bare er feil, altså ren støy.

7.1.1 Homonymi

Homonymi er en generell utfordring i emneordsarbeid, og var også en utfordring i forbindelse med mapping. Spesielt i tilfeller hvor det ikke var knyttet noen annen informasjon til emneordene, som i eksempelet med Antenner i Realfagstermer (Kuldevere et al., 2014). Uten en hierarkisk struktur, tilleggskommentar eller tilknytning til et klassenummer var det nødvendig å innhente informasjon fra katalogen, altså konteksten emneordet var benyttet i, for å avgjøre hva slags antenner det var snakk om.

I dataanalysen var det ikke mulig å umiddelbart identifisere noen utfordringer knyttet til homonymi. Det er mulig at en automatisk sammenkobling av de to vokabularene ville avdekket noen homonymer, men ved den manuelle metoden for identifisering som ble benyttet i dataanalysen var det altså ikke mulig å oppdrive noen eksempler. Dette kan tyde på at emneord i strenger reduserer muligheten for misforståelser ved homonymi fordi emneordene settes i en tydelig sammenheng som uttrykker konteksten de kan benyttes i.

Mangelen på homonymi kan også være et resultat av vokabularenes bruksområder og bakgrunn. Mens Humord benyttes av fagbibliotek, benyttes BIBBI hovedsakelig til indeksering for folkebibliotek, og består dermed muligens av mer allmenne begrep, mens Humord inneholder fagbegrep. Begge vokabularene er omfangsrike og godt etablerte, og det legges mye arbeid i dem. Dermed kan det hende at det er en god homonymikontroll innbyrdes i hvert system, hvor man heller velger et synonym som ikke har homonymer som foretrukken term, eller formulerer ordet på en måte som gjør at man kan unngå homonymi. For eksempel kunne man valgt, dersom det beskrev alle typer antenner man ønsker å beskrive i Realfagstermer, å kalle det Parabolantenner i stedet for bare Antenner.

I forbindelse med mappingen av Realfagstermer og TEKORD ble det foreslått å innføre kvalifikatorer for å begrense utfordringer knyttet til homonymi (Kuldevere et al., 2013, s. 16). For å etablere interoperabilitet for området hvor Realfagstermer og TEKORD overlapper, er dette en overkommelig oppgave da det ikke utgjør så mange termer. I en større sammenheng, skulle man for eksempel ønske en videre mapping mot et tredje vokabular, kan det fort virke mot sin hensikt. For å kunne mappe emneord som har homonymer mot rett begrep i et annet vokabular og ikke et homonym, må man nødvendigvis også ha samme kvalifikator dersom man ønsker å benytte seg av automatisk mapping basert på ordlikhet. Likevel kan et bevisst forhold og arbeid med homonymi internt i hvert vokabular være gunstig for å unngå utfordringer med homonymi ved mapping.

7.1.2 Sammensatte begreper

Gödert, Hubrich & Nagelschmidt skriver at ved mapping av sammensatte begreper som emneord i streng må meningen i både termene og relasjonene mellom dem kunne forstås (2014, s. 134-136). Emneord i streng er helt klart mer utfordrende å mappe i sin helhet enn emneord basert på postkoordinering. I de to prosjektene hvor man mappet Realfagstermer, ble strenger ganske enkelt utelatt fra forsøkene fordi man ikke fikk det til å fungere. Det foreslås likevel at emneord i streng kan være lettere å mappe til Dewey fordi de begge baserer seg på prekoordinering (Kuldevere et al., 2014, s. 19).

Den muligheten Dewey tilbyr for bygging av numre byr på utfordringer når man mapper emneord basert på postkoordinering. Et godt eksempel på dette er begrepet terapi, som ble diskutert i både dokument- og dataanalysen. I mappingen av Realfagstermer mot WebDewey oppstod det utfordringer knyttet til begrepet fordi det er et emneord man gjerne benytter i kombinasjon med den tilstanden som behandles (Kuldevere et al., 2014, s. 12). I Dewey og i BIBBI, som baserer seg i stor grad på Dewey, benyttes begrepet også som et slags suffiks, hvor man legger terapi etter tilstanden som behandles. Dermed blir "terapi" underordnet en rekke sykdommer. I dataanalysen ble det vist hvordan det enkeltstående emneordet Terapi i BIBBI dermed ble et slags samle-emneord for dokumenter som var for spesielle til å få et eget emneord, slik at en mapping mellom Terapi i BIBBI og Terapi i Humord ville vært svært lite fruktbart.

I MACS oppstod det også utfordringer knyttet til at de forskjellige språkene i større og mindre grad tillater orddeling. Norsk språk er, som tysk, tilbøyelig til å dele opp færrest mulig ord. Dette aspektet ved sammensatte begreper kunne ikke undersøkes videre i dataanalysen da den bare forholdt seg til norske emneord, men det er en faktor som er verdt å tenke over ved mapping mot for eksempel engelske vokabularer. Her kan sammensatt ekvivalensmapping som beskrives i standarden være en mulig løsning (ISO 25964-2, 2013, s. 22-24)

7.1.3 Hierarki

Innen kategorien hierarki kan man igjen si at det er to typer utfordringer. Den ene er utfordringen som eksemplifiseres i Gödert, Hubrich & Nagelschmidt hvor man i SWD bare forholdt seg til Papageien, mens man i Dewey omtaler papegøyer i flere kontekster, både biologisk sett og som kjæledyr (s. 109-111). Denne typen utfordring så vi igjen i dokumentanalysen hvor for eksempel Knokler ble behandlet uavhengig av hvorvidt det var knokler hos dyr eller mennesker i Realfagstermer, mens det i Dewey er et skille mellom de to (Kuldevere et al., 2014). Denne utfordringen knyttet til hierarki vil være spesielt tydelig ved mapping mellom et emneordssystem og et klassifikasjonssystem, ettersom emneordssystemer sorteres etter emne, mens klassifikasjonssystemer sorteres etter fag. Denne problematikken ble også nevnt i forbindelse med mapping av Humord til DDC, eksemplifisert med emneordet Døden (Gulbrandsen, Heggø, Knutsen & Seland, 2015, s. 11).

I mappingen av Realfagstermer mot WebDewey ble det bare mappet mot deler av Dewey. Det oppstod flere tilfeller av emneord som passet inn flere steder i hierarkiet, men det ble også antatt at man ved mapping mot hele Dewey ville få enda flere slike tilfeller (Kuldevere et al., 2014). Spesielt ved mapping av vokabularer som er beregnet på et begrenset fagområde, som for eksempel Realfagstermer er, kan en slik begrensning være svært ressursbesparende samtidig som det ikke går spesielt mye på bekostning av gjenfinningseffektiviteten.

Det er helt klart mulig at man risikerer å senke fullstendigheten noe ved å ikke mappe til alle potensielle matcher, men samtidig kan presisjonen forbedres. En slik begrensning foreslås også i ISO-standarden. I eksempelet med Samer i dataanalysen ble det illustrert hvordan noen begreper kan være spredt over flere fagområder. Her kan en slik begrensning være mindre gunstig. Totalt

sett må man gjøre en vurdering av ressursbruk og hva man ønsker å oppnå med mappingen før man tar et slikt valg.

Den andre utfordringen innen hierarki er også knyttet til bruk av emneordsvokabularene, og dreier seg om forskjellig grad av spesifisitet, som i eksempelet med forskjellige typer ugler i Gödert, Hubrich og Nagelschmidt (2014, s. 115-116). Som nevnt kan dette noen ganger være en fordel fordi man får samlet "smale" emneord. I dataanalysen ble et slikt tilfelle eksemplifisert gjennom begrepene "Feministisk økonomi" og "Økonomi og feminisme" Ved direkte mapping mellom BIBBI og Humord var dette relativt uproblematisk da begge vokabularene hadde passende emneord. I Dewey falt begge begrepene innunder samme klassenummer.

Selv om de to begrepene i utganspunktet beskriver et forskningsfelt og det som forskes på, er det ikke utenkelig at en indekserer anser "Økonomi og feminisme", enten kombinert ved prekoordinering som i BIBBI eller ved postkoordinering som det er rom for i Humord, som tilstrekkelig spesifikt for dokumenter om feministisk økonomi. Det er for så vidt heller ikke direkte feil å benytte begge begrepene som emneord. Man kan også tenke seg at noen velger å henvise mellom begrepene med en Se også-henvisning.

7.1.4 Kontekst og forskjellig bruk

Også her kan man skille mellom to hovedtrekk ved utfordringene. Den ene utfordringen er knyttet til avvik fra standarden som settes for struktur og bruk av de forskjellige vokabularene. Som illustrert med "Library", "Public Library" og "Documentation center" av Gödert, Hubrich & Nagelschmidt (s. 107-108), valgte man i den ene tesaurusen å etablere "Brukt for"-henvisninger for "Public Library". Dette er egentlig en underordnet term med en generisk relasjon til "Library". "Documentation" burde egentlig fått en "Se også"-henvisning i kraft av sin assosiative relasjon til "Library." i ett system er denne praksisen helt uproblematisk, og å gi alle termer helt korrekte henvisninger og hierarkiske plasseringer ville vært unødvendig arbeid dersom man fint kan indeksere alle dokumentene i samlingen med "Library" og fortsatt ivareta tilstrekkelig presisjon for bibliotekets brukergruppe. Ved mapping kan det derimot by på utfordringer.

I mappingen av Realfagstermer og TEKORD viste det seg at en slik praksis bød på utfordringer Innbyrdes da man hadde satt SE-henvisninger fra blant annet Demninger til Dammer i det ene vokabularet, mens Demninger var en foretrukket term i det andre vokabularet. I dette tilfellet hvor man benyttet automatisk genererte mappinger ble da Dammer foreslått som synonym til Demninger (Kuldevere et al., 2013).

Et lignende tilfelle ble diskutert i dataanalysen tilknyttet demens. Her er Senil demens foretrukket term i Humord, med Brukt for-henvisning til både Alzheimer og Demens. I Dewey og BIBBI er Alzheimer sidestilt med Demens mens Senil demens er plassert under geriatri. Selv med en-til-mange-mapping og flere typer mappingrelasjoner i bruk vil det her være vanskelig å mappe Humords demens-begrep til alle BIBBIS demens-begrep da konteksten er så forskjellig. I BIBBI har man etablert skiller mellom de forskjellige stadiene og typene av demens, mens man i Humord har samlet alt under ett og dermed ikke har mulighet for å spore konteksten på samme måte.

Den andre hovedtendensen innunder kategorien som dreier seg om forskjellig bruk er kontekst, som har vist seg å være en svært avgjørende faktor i alle mappinger. Emneordenes kontekst innebærer blant annet at man aldri vil være i stand til å mappe nøyaktig det samme begrepet fordi konteksten det ene emneordet benyttes i og meningen som knyttes til det vil aldri kunne gjenskapes i et annet tilfelle. Heldigvis vil meningsforskjellen ved de fleste begreper være tilstrekkelig liten til at man i gjenfinningsforemål ikke får noen problemer ved å mappe disse begrepene. I noen tilfeller kan det faktisk være en fordel å mappe begreper som er oppstått i forskjellige kontekster fordi det vil øke tilfanget av tolkninger av et dokuments innehold. Dersom alle disse tolkningene kan bli ansett som relevante, blir også fullstendigheten større.

Gödert, Hubrich & Nagelschmidt (s. 107-108) skiller mellom mapping hvor man tar hensyn til emneordenes kontekst og mapping hvor man ikke gjør dette. Basert på antagelsen om at de fleste indekseringsspråk struktureres og benyttes nesten likt er det mulig å opprette en rekke gode mappinger ved å se bort fra emneordenes kontekst, men resultatene fra analysen viser likevel at en slik tilnærming ikke er uproblematisk.

7.1.5 Kategorienes tilstrekkelighet

Delvis i dokumentanalysen og langt mer tydelig i dataanalysen var flere av eksemplene passende i flere av kategoriene. For eksempel kunne eksempelet med terapi både kunne sies å være en utfordring knyttet til hierarkisk struktur og til måter å uttrykke sammensatte begreper på. Dette kan tyde på flere ting, blant annet at kategoriene ikke er tilstrekkelig begrenset og definert. Det kan også tyde på at mapping, og spesielt mappingen av to så forskjellige vokabularer som BIBBI og Humord, er et mer sammensatt arbeid enn det kan fremstå ved et førsteinntrykk. I videre arbeid kan det være verdt å vurdere hvorvidt nærsynonymi og polysemi bør innføres som kategori. Problematikk knyttet til dette trekkes fram blant annet i forbindelse med hierarki og bruk, i form av valg av se-henvisning i stedet for ”se også”, og hva dette gjør med funn av synonymer ved mapping.

Det kan også være aktuelt å inndele hierarki-kategorien i to deler, hvor den ene omhandler forskjellig hierarkisk plassering av emner, et fenomen som typisk vil oppstå ved mapping mellom et emneordssystem og et indekseringsspråk. Dette oppstår også i stor grad ved mapping mellom Humord og BIBBI ettersom BIBBI kan sies å ha hierarkisk struktur i Dewey. Den andre delen av hierarki-kategorien er utfordringer knyttet til ulik grad av spesifitet. Et vokabular kan ha fordelt underkategorier på flere emner der det andre har samlet alle disse underkategoriene under en overordnet term.

Kategorien for ulik bruk og kontekst kan eventuelt også deles i to kategorier, hvor den ene inkluderer utfordringer som oppstår som en konsekvens av at man internt i de enkelte vokabularene har lempet litt på reglene og for eksempel satt bruk-henvisninger for termer som egentlig ikke er synonymer, men som står i en annen relasjon til den foretrukne termen. Som i eksempelet fra dataanalysen med Demens i Humord, hvor man hadde benyttet Bruk på både overordnede, assosiative og sideordnede relasjoner, sammen med synonymer, ser man hvilke utfordringer interne praksiser kan skape når man forsøker å kombinere dem med andre systemer.

Det kan også være nyttig å etablere utfordringer knyttet til emneordenes kontekst som en egen kategori. Analysen i denne oppgaven har ikke viet så stor oppmerksomhet til hvilke dokumenter som faktisk blir mappet. En slik undersøkelse ville vært enklere å gjennomføre på et ferdig

mappet vokabular. Likevel er det flere indikasjoner på at kontekst kan gi emneordene litt forskjellig mening.

Det kan også vurderes hvorvidt kategoriene knyttet til ressursbruk og konsistens ved valg av relasjoner er nødvendige å beholde som kategorier. Ettersom de er knyttet til selve mappingarbeidet fremfor indekseringsspråkenes struktur må man forholde seg til dem på et litt annet plan. Spesielt kategorien som indikerer at mappingarbeid er svært ressurskrevende er dessuten relativt selvsagt. På en annen side er det godt mulig at overveielser knyttet til strukturen i vokabularene man skal mappe, og fremgangsmåten og mappingmetodene man velger, vil ha innvirkning på disse problemene.

7.2 Målet og midlene

Formålet med mapping, som vi har sett nevnt mange ganger, er å samordne vokabularer. Dette har igjen to effekter: indekserer kan få tilgang til flere forslag til emneord, synonymer og lignende ved indeksering og man kan i søk åpne for muligheten til å søke på tvers av flere samlinger indeksert med forskjellige vokabularer.

Et felles mål for alle prosjektene som ble omtalt i dataanalysen, og for mapping generelt, er å samordne. I MACS var det et mål å etablere muligheter for søk på tvers av flere språk. Søk på tvers av kataloger var også et mål i CrissCross, sammen med å forbedre søkemulighetene som finnes blant annet ved å gi muligheter for rangering. I mappingen av Realfagstermer til TEKORD og til WebDewey var ikke mappingen målet i seg selv, prosjektene var derimot oppstartet for å undersøke metodikk for mapping. For mappingen av Realfagstermer og TEKORD var det et mål å berike vokabularene, blant annet ved å hente inn synonymer fra hverandres vokabular. I mappingen av Realfagstermer mot DDC var spesielt automatikk i fokus, og det ble blant annet undersøkt hvorvidt en automatisk kobling mellom vokabularene kunne være et utgangspunkt for mapping og lette det manuelle arbeidet.

Skal man oppnå disse målene, og spesielt søk på tvers av baser, er det en nødvendighet å vurdere hvordan man kan ivareta både presisjon og fullstendighet på best mulig måte, samtidig som arbeidsmengden ikke blir uoverkommelig. På en side vil man ivareta presisjonen best mulig ved

å bare mappe begreper som ligner tilstrekkelig mye, og på en annen side vil man ivareta fullstendigheten ved å ikke utelukke mappinger. Her spiller de forskjellige mappingrelasjonene inn. I tillegg til relasjonene som presenteres i ISO-standarden, har også de tyske "degrees of deteminancy" i CrissCross-prosjektet blitt nevnt. Disse 4-5 relasjonene er langt færre enn de 19 relasjonene som ble foreslått av Chaplan i 1995. Trolig er det tilstrekkelig med relasjonene fra standarden eller CrissCross for å kunne ivareta presisjonen i søk.

7.2.1 Valg av relasjoner

Ved å skille mellom grader av likhet kan man også rangere trefflister ved hjelp av mappingrelasjonene som er opprettet. Dette er felles for ISO-standardens relasjoner og "degrees of determinacy". ISO-standardens relasjoner er sterkt preget av at standarden er for interoperabilitet mellom en tesaurus og andre vokabularer, og har dermed en stor likhet med relasjonene som benyttes internt i en tesaurus. "Degrees of determinacy" er konstruert spesielt med tanke på mapping mot DDC. Det kan derfor tenkes at de kan være mer egnet ved mapping av for eksempel en flat struktur mot DDC. Likevel vil det, med tanke på interoperabilitet i det store bildet, sannsynligvis være gunstig å rette seg etter ISO-standarden selv om man mapper med noe annet enn en tesaurus. McChulloch & MacGregor anbefalte i sin vurdering av Chaplans 19 relasjoner at mappingrelasjoner må kunne anvendes på alle typer kunnskapsorganisasjonssystemer (2008).

I dataanalysen er det i utgangspunktet ikke tatt stilling til hvilke relasjonstyper som bør velges i de forskjellige tilfellene, selv om noen løsninger foreslås. Formålet med dataanalysen var å undersøke hvorvidt kategoriene som ble identifisert i dokumentanalysen var av en allmenn karakter, gjennom å peke på grunnleggende forskjeller i forskjellige indekseringsspråk, med BIBBI, Humord og DDK5 som eksempler. Spesielt utfordringer knyttet til hierarki kan muligens løses ved å velge de hierarki-baserte BM og NM fra ISO-standarden, men det vil ikke løse alle utfordringer.

I dokumentanalysen ble det avdekket at flere syntes det var utfordrende og ressurskrevende å velge relasjonstype. Disse utfordringene ble grunnlaget for en av de sju kategoriene. Utfordringene kan reduseres noe ved å etablere rutiner for å sikre best mulig vurdering av

relasjonene, for eksempel ved at to fagpersoner vurderer alle relasjoner, eller i det minste relasjoner hvor den første som vurderer er i tvil.

Selv med en slik kontroll kan man vanskelig sikre at alle personer som vurderer relasjoner tenker likt i tvilstilfeller som ligner hverandre. Så lenge menneskelig vurdering er en faktor, vil det også være mulighet for redusert konsistens. Dersom denne typen utfordringer vurderes som alt for stor, kan det være en mulighet å vurdere om man behøver å benytte alle relasjonstypene oppgitt i ISO-standarden for å oppnå målene man setter for mappingen. En mapping med for eksempel bare ekvivalens- og tilnærmet ekvivalensrelasjoner vil fortsatt åpne for relativt mange søk på tvers av baser.

7.2.2 Prekoordinerte og postkoordinerte utgangspunkt

Man skulle kanskje tro at prekoordinerte emneord og emneord basert på postkoordinering er to uforenelige fenomener som aldri kan mappes med hverandre. Delvis kan man si at dette stemmer, i hvert fall hvis man skal beregne en rimelig mengde ressursbruk på arbeidet. Mens prekoordinerte vokabularer har muligheten for å uttrykke sammensatte begreper, og dermed uttrykke konteksten de står i mer tydelig, er postkoordinerte emneord langt mer fleksible. De er enklere å lage og vedlikeholde, og er langt mer tilpasset søkemotorer og semantisk web (Ohren, Rydland & Rype, 2013, s. 16). Nettopp dette med semantisk web er et viktig poeng i forbindelse med mapping. Mapping og linked data står i samme tanketog om å tilgjengeliggjøre og dele data.

Det har blitt foreslått at prekoordinerte emneord kan være enklere å mappe mot Dewey (Kuldevere et al., 2014, s. 19). Denne påstanden har ikke blitt undersøkt i denne oppgaven, da BIBBI allerede er knyttet opp mot DDK5. Med dette utgangspunktet vil det være lettere å eventuelt mappe BIBBI mot WebDewey, og på den måten få en tilknytning til en rekke andre vokabularer. Samtidig oppstår en slags mapping mellom WebDewey og DDK5 som kan være nyttig i situasjoner hvor man har behov for å sammenligne klassenumre.

Emneord i streng kan tenkes å kunne mappes med en mindre grad av utfordringer knyttet til sammensatte begreper, dersom det mappes til et annet indekseringsspråk basert på prekoordinering. Dette forutsetter at man har kombinert og benyttet begreper på samme måte i

begge vokabularene. I dataanalysen ble det også antydet at mapping med prekoordinerte vokabularer ikke vil ha en like stor grad av tilfeller av homonymi som postkoordinerte vokabularer eller mapping med prekoordinerte vokabularer hvor man velger å splitte strengene. Vokabularer basert på postkoordinering, derimot, virker mer fleksible og vil trolig ha færre utfordringer knyttet til hierarki. Totalt sett virker postkoordinerte vokabularer mer egnet for mapping basert på automatiske koblinger mellom emneordssystemene.

8 Oppsummering

Denne oppgaven har forsøkt å besvare følgende problemstilling:

Hva slags utfordringer kan oppstå ved forskjellige mappinger av indekseringsspråk?

- Hvilke fremgangsmåter er benyttet og hvilke utfordringer har oppstått i tidligere prosjekter?
- Hvilke fordeler og ulemper innebærer de forskjellige fremgangsmåtene?
- Hvilke fordeler og ulemper innebærer mapping av forskjellige typer indekseringsspråk?

For å besvare problemstillingen ble det benyttet en todelt metode. Først ble det foretatt en dokumentanalyse hvor det ble identifisert sju kategorier av utfordringer:

- Homonymi
- Forskjellige hierarkiske plasseringer eller grader av inndeling
- Støy
- Sammensatte begreper
- Forskjellige begreper som et resultat av forskjellig bruk og kontekst ved indeksering
- Konsistens i forbindelse med menneskelig vurdering av relasjonene
- Ressursbruk

I forkant av dokumentanalysen ble også de fire prosjektene som var representert i dokumentanalysen presentert i kapittel 4 «Om prosjektene». Her ble også målene og fremgangsmåten i de forskjellige prosjektene diskutert. Målene og fremgangsmåtene er relativt like, men det er blitt valgt noe forskjellige grader av automatisering og også forskjellige relasjonstyper for mappingrelasjonene. Ingen av disse valgene virket å ha stor innvirkning på kategoriene av utfordringer som ble presentert, da de fleste kategoriene ble identifisert gjennom dokumentanalyse i alle prosjektene.

I dataanalysen ble kategoriene som kan knyttes til indekseringsspråks struktur og bruk anvendt på to nye sett med indekseringsspråk: Humord og BIBBI, både i eksempler hvor man antok at vokabularene ble direkte mappet, og i eksempler hvor man antok at de ble mappet via klassenumre i DDK5. Her ble det funnet at utfordringene var av en allmenn karakter, med unntak

av homonymi, som ikke kunne gjenskapes med de to systemene. Dette er trolig en større utfordring ved vokabularer basert på enkeltstående emneord, eller dersom man velger å splitte strenger.

I diskusjonen ble funnene fra analysen diskutert i lys av teorien, hvor det er en viss grad samsvar med tidligere uttalelser i forskningslitteraturen om potensielle utfordringer. Der ble det foreslått en mulig videre underinndeling av kategoriene samt noen nye kategorier, blant annet ble det vurdert hvorvidt utfordringer knyttet til nærsynonymer burde være en egen kategori. Det ble også diskutert om det er mulig å bare benytte enkelte typer mappingrelasjoner for å begrense utfordringer knyttet til konsistens ved valg av relasjoner, samt ressursbruk.

Videre ble det diskutert i hvor stor grad de forskjellige utfordringene er tilstede når man mapper med henholdsvis postkoordinerte og prekoordinerte indekseringsspråk. Tilsynelatende vil det være mindre utfordrende å mappe prekoordinerte emner til et klassifikasjonssystem som tillater nummerbygging, som for eksempel Dewey som det er mappet mye til. Postkoordinerte emneord, på sin side, fremstår som mer fleksible og vil trolig ha færre utfordringer knyttet til hierarkisk plassering.

8.1 Metodiske svakheter og videre forskning

Det er viktig å understreke at hele denne oppgaven er svært problemorientert. Det er blitt viet lite plass til de positive effektene ved mapping, og alle emneordene som lett og uproblematisk lar seg kombinere med emneord fra et annet vokabular er ikke viet noen oppmerksomhet i det hele tatt. Det er tross alt disse emneordene som utgjør flertallet ved de fleste mappinger. Oppgaven har heller ikke diskutert alle aspekter ved mapping. På denne måten sier den mer om grunnleggende forskjeller mellom emneordssystemer enn den sier om mappingarbeidet som en helhet.

I svært mange mappingprosjekter, for eksempel mappingen av Realfagstermer til TEKORD og Realfagstermer til WebDewey, struktureres emneordene i et linked data-passende format. Linked data og mapping av emneord er derfor tett knyttet sammen. Likevel er det ikke viet mye oppmerksomhet til linked data i denne oppgaven. Som nevnt i forbindelse med FinnOnto i teorikapittelet er det ikke nødvendigvis uproblematisk å strukturere for eksempel en tesaurus som

linked data. Her er kjernen av problemet mye det samme som i kategoriene som er identifisert i denne oppgaven - at mye av strukturen man legger i indekseringsspråk gir mening for mennesker, men er uleselig for en datamaskin.

Begrensningen som er gjort her er først og fremst et valg som er tatt med hensyn til plass- og ressursbruk. En fordypning i disse problemstillingene ville økt omfanget på oppgaven betraktelig, riktig nok med svært relevant informasjon og poenger, men likevel ville tematikken ikke fått oppmerksomheten den fortjener. Det er derfor høyst relevant å videreføre tankegangen som er presentert i denne oppgaven i forhold til emneords mening og kontekst, og meningen i relasjonene mellom emneordene, og hvordan man kan ivareta dette i en linked data-tilpasset struktur. Videre vil det også være interessant å vurdere hvordan mappede emneordssystemer kan berike og supplere, samt utnytte, andre typer metadata tilgjengelig som linked data. For eksempel i forbindelse med nyere arbeid og tankeganger innen katalogisering.

Også de forskjellige måtene å uttrykke relasjonene mellom begreper som mappes kan godt vurderes nærmere. Her er det allerede gjort en del arbeid med utgangspunkt i Chaplans 19 relasjonstyper, og valget av relasjonstyper er allerede satt av ISO-standard 25964-2 i mappingarbeid hvor ett av vokabularene er en tesaurus. Hvordan disse relasjonene forholder seg til andre relasjonstyper, skulle man forsøke å koble sammen flere mappinger, er verdt videre undersøkelser. En vurdering av hvorvidt relasjonstypene fra ISO-standarden er gode nok for mapping av andre indekseringsspråk enn tesaurus er også interessant.

To av kategoriene som ble identifisert er knyttet til at mappingarbeid er krevende, både i form av ressursbruk, og også i form av å vurdere og velge relasjonstyper på en konsistent måte. I den forbindelse kan det være nyttig å undersøke om man kan oppnå et tilfredsstillende resultat ved mapping med et begrenset utvalg av mappingrelasjonene fra standarden, for eksempel kun ekvivalens og tilnærmet ekvivalens. Det har også blitt foreslått om utfordringer knyttet til hierarkisk plassering kan begrenses noe ved å bare mappe til deler av målvokabularet, spesielt i tilfeller hvor kildevokabularet er begrenset til et visst fagområde. Hvilke konsekvenser dette har for fullstendighet og presisjon, samt hvor arbeidsbesparende det kan være, kan også undersøkes.

I dataanalysen ble DDK5 valgt som datagrunnlag for tilfeller hvor man vurderte klassenumre fra Dewey. Dette er i utgangspunktet ikke representativt da de fleste mappingarbeider i dag mapper mot den fullstendige versjonen. Med tanke på dette er ikke dataanalysen en tydelig indikator på hvordan en mapping av både BIBBI og Humord til Dewey vil utarte seg, og hvordan vokabularene vil forholde seg til hverandre i en slik situasjon. Det er heller ikke meningen. Når den fullstendige norske oversettelsen av Dewey publiseres, og begge vokabularene eventuelt blir mappet til dette, vil en ny test av kategoriene på disse nye dataene være interessant.

Innledningsvis ble samordning nevnt som en sentral motivasjon for mapping. I oppgaven har en rekke faktorer som kan forbedre eller forhindre samordning blitt diskutert, men de faktiske effektene av mapping har ikke blitt viet så mye oppmerksomhet som det kanskje fortjener. Spesielt ved søk på tvers av kataloger bør det gjøres undersøkelser av hvordan de forskjellige kategoriene spiller inn på fullstendighet og presisjon. Dette gjelder både kategoriene som ble diskutert i dataanalysen, men også ressursbruk, støy og konsistens ved valg av relasjoner, som også vil spille inn på fullstendighet og presisjon.

Problemstillingen etterspør utfordringer som kan oppstå ved mapping av indekseringsspråk. De vanligste typene indekseringsspråk og måtene å organisere disse på er representert i oppgaven. Dette inkluderer vokabularer med flat struktur og tesaurusstruktur, prekoordinerte og postkoordinerte emneord, klassifikasjonsskjema med mulighet for nummerbygging, og vokabularer organisert etter fag og emne. Dette dekker likevel ikke alle måter å organisere emneord på, og utfordringene som er identifisert vil ikke nødvendigvis være like store i for eksempel ontologier. Om kategoriene gjelder også her, er verdt nærmere undersøkelser.

9 Litteraturliste

Braun, V., & Clarke, Victoria. (2013). *Successful qualitative research : A practical guide for beginners*. Los Angeles, Calif: Sage.

Chaplan, M. A. (1995). Mapping Laborline thesaurus terms to Library of Congress subject headings: implications for vocabulary switching. *Library Quarterly* (65). doi: 10.1086/602752

Gulbrandsen, A., Heggø, D. M. O., Knutsen, U. & Seland, G. (2015) *På vei mot en generell norsk tesaurus?: delprosjekt Metodikk for mapping av Humord mot WebDewey: Rapport.* Oslo: Universitetsbiblioteket i Oslo. Hentet fra http://www.ub.uio.no/for-ansatte/om-ubo/prosjekter/tesaurus-mapping/delte-dokumenter/prosjektrapport-tesaurus-mapping.pdf

Gödert, W., Hubrich, J, & Nagelschmidt, M. (2014). *Semantic knowledge representation for information retrieval*. Berlin: Walter De Gruyter.

Hjortsæter, E. (2009). Emneordskatalogisering: Innholdsanalyse, emnerepresentasjon og lagring (3. utg.). Oslo: ABM-media.

Hougaard, B. S. (2011, 11. juli). *Om tesaurusen Humord.* Hentet 13. mai 2015 fra http://www.bibsys.no/files/out/humord/om-tesaurusen.html

Hubrich, J. (2010). Intersystem relations: characteristics and functionalities. I: *Concepts in Context. Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization July 19th - 20th, 2010.* (s.37 - 50). Würzburg: Ergon, 2011

Hyvönen, E., Viljanen, K., Tuominen, J. Kauppinen, T., Ruotsalo, T., Valkeapää, O., ...Kurki, J. (2007). *Elements of a National Semantic Web Infrastructure: Case Study Finland on the Semantic Web.* Helsinki: Helsinki University of Technology. Hentet frahttp://www.seco.tkk.fi/publications/2007/hyvonen-et-al-elements-2007.pdf

International Organization for Standardization ISO 25964-2. Information and documentation : Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies 2013, Geneve: ISO.

Jacobs, J. H., Mengel, T. & Müller, K. (2010) Insights and Outlooks: A Retrospective View on the CrissCross Project. I: *Concepts in Context. Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization July 19th - 20th, 2010.* (s.37 - 50). Würzburg: Ergon, 2011

Jahns Y. & Karg, H. (2010) Translingual retrieval: moving between vocabularies - MACS 2010. I: *Concepts in Context. Proceedings of the Cologne Conference on Interoperability and Semantics in Knowledge Organization July 19th - 20th, 2010.* Würzburg: Ergon, 2011

Johannessen, A., Tufte, P. A. & Christoffersen, L. (2010) Introduksjon til samfunnsvitenskapelig metode. Oslo: Abstrakt.

Kuldvere, V., Lundevall, M., Hegna, K., Konestabo, H. S., Låberg, K. T., Flatby, E. S., Greenall, R. (2013) *Realfagstermer og TEKORD: RDF som plattform for sammenlikning ogsammenføyning av emnesystemer?*.Oslo:Universitetet i Oslo. Hentet fra https://www.ub.uio.no/om/prosjekter/avsluttet/realfagstermer-tekord/realfagstermer-og-tekord-rapport.pdf

Kuldvere, V., Flatby, E. S.,Heggø, D. M. O., Konestabo, H. S., Lundevall, M. & Låberg, K. T. (2014). *Felles terminologi for klassifikasjon med Dewey.* Oslo: Universitetet i Oslo. Hentet fra http://www.ub.uio.no/om/prosjekter/deweymapping/prosjektrapportfelles-klassifikasjon.pdf

McCulloch, E. & MacGregor, G. (2008). Analysis of equivalence mapping for terminology services. *Journal of Information Science,* (34) doi: 10.1177/0165551507079130

McCulloch, E. & Nicholson, A. S. D. (2005).Challenges and issues in terminology mapping: a digital library perspective. *The Electronic Library*, (23), 6. doi: 10.1108/02640470510635755

Ohren, O. P., Heggø, D. M., Hougaard, B. S. Johnsen, L., Knutsen, U.Kuldevere, L. V., …Tjelta, Torstein. (2015) *En norsk generell tesaurus? Sluttrapport med anbefalinger fra Tesaurus forprosjekt (5.3.2014-1.3. 2015).* Oslo: Nasjonalbiblioteket. Hentet fra http://www.nb.no/content/download/10372/98161/file/Tesaurus-forprosjekt-rapport-1.0.pdf

Ohren, O. P., Rydland, K. & Rype I. (2013). *Emneinnganger i Nasjonalbiblioteket: Anbefalinger om praksis for emnebeskrivelse: Del 1: Materiale med verbalt innhold.* Oslo: Nasjonalbiblioteket. Hentet fra http://www.nb.no/content/download/8072/80266/file/Anbefaling-emnebeskrivelse-NB-v1.0.pdf

Svanberg, M. (2006) *Övergång till Dewey Decimal Classification. Vad skulle det innebära?: Delstudie 3 i Katalogutredningen.* Stockholm: Kungliga biblioteket. Hentet fra http://www.kb.se/Dokument/Om/projekt/avslutade/katalogutredning/delst3_slutrapport.pdf

Vagle, W., Sandvik, M. & Svennevig, J.. (1993). Tekst og kontekst : En innføring i tekstlingvistikk og pragmatikk (Vol. Nr 73, Skriftserie (Landslaget for norskundervisning.)). Oslo: Cappelen.

Annet:

Data fra Humord er hentet fra http://wgate.bibsys.no/search/pub?base=HUMORD mars-juni 2015

Data fra Biblioteksentralen er hentet via Promus mars-juni 2015