Felles terminologi for klassifikasjon med Dewey

Rapport

Viola Kuldvere, Ellen Samdahl Flatby, Dan Michael Olsen Heggø,
Heidi Sjursen Konestabo, Mari Lundevall, Kyrre Traavik Låberg

01.07.2014

Innhold

1 Innledning 3
1.1 Bakgrunn 3
1.2 Mål 4
1.3 Prosjektgruppen 4
2 Datagrunnlag 5
2.1 TEKORD 5
2.2 Realfagstermer 5
2.3 TORT 5
2.4 DDC23 5
2.5 Katalogdata 6
3 Utvikling av metoder 6
3.1 Mapping 6
3.1.1 Definisjoner 6
3.2 Formelle avgrensinger 7
3.3 Automatiske mappeforslag 7
3.3.1 Tekstsammenlikning 8
3.3.2 Statistisk sammenlikning 8
3.4 Intellektuelle vurderinger 10
4 Resultater og diskusjon 10
4.1 Forslag til mapping 10
4.2 Verktøy 13
4.2.1 Bakgrunn: Regneark 13
4.2.2 µmapper: Datamodellen 14
4.2.3 µmapper: Applikasjonen 16
4.3 Arbeidsbeskrivelse for mapping 17
4.4 Realfagstermer og TEKORD 18
4.4.1 Kvalitetssikring av eksisterende data 18
4.4.2 Videre samarbeid 19
4.5 Generell diskusjon 19
4.5.1 Sammenlikning av tekststrenger – metoden 19
4.5.2 Hypotese 20
4.5.3 Mapping basert på statistisk sammenlikning 20
4.5.4 Vurdering av relasjonstyper 20
4.5.5 Overganger mellom UDC og DDC 21

5 Oppsummering og konklusjon 21
Arbeidsbeskrivelse for mapping Realfagstermer/TEKORD/DDC 23
Referanser 25

1 Innledning

1.1 Bakgrunn

Prosjektet *Felles terminologi for klassifikasjon med Dewey* ved Realfagsbiblioteket, Universitetsbiblioteket i Oslo (UBO) ble tildelt samarbeids- og utviklingsmidler fra Nasjonalbiblioteket for 2013. Prosjektet har samarbeidet med Ellen M. S. Flatby fra NTNU Universitetsbiblioteket (NTNU UB).

Prosjektet har hatt som mål å utvikle metoder for mapping (kobling) fra emneordssystemene Realfagstermer (RT)[1] og TEKORD (TO) til Deweys desimalklassifikasjon (DDC). Dette er en videreføring av prosjektet *Realfagstermer og TEKORD* [1] som brukte RDF/SKOS som plattform for sammenlikning av terminologi i Realfagstermer og TEKORD. I prosjektet testet vi en automatisk metode som gikk ut på å sammenlikne termers tekstlikhet for å finne en overlapp mellom vokabularene med tanke på mulig berikelse, samordning og utveksling av emneord. Metoden fant i overkant av 3000 overlappende begreper, noe som utgjorde rundt 20 % av hvert vokabular. Denne overlappen av begreper i de to emnesystemene utgjør nå en viktig del av datagrunnlaget for mapping mot DDC.

Begge prosjektene har sitt utspring i rapporten *Bibliografisk og emnemessig beskrivelse av UBOs samlinger* [2]. Rapporten peker på den store variasjonen i emneord og klassifikasjon ved UBO. Det brukes mange forskjellige klassifikasjonssystemer, og emneordspraksisen er ulik. Samordning og standardisering av systemene og praksisen er nødvendig for at disse systemene skal kunne «kommunisere» med andre systemer nasjonalt og internasjonalt og slik også lette gjenbruk av registreringsdata og gi god gjenfinningskvalitet for informasjonssøkeren.

I internasjonal sammenheng er arbeidet med samordning, standardisering, overganger mellom ulike emne- og klassifikasjonssystemer, samt utprøving av semantiske web-teknologier, som lenkede data, kommet lenger enn i Norge. Som eksempler kan nevnes mapping mellom Svenska ämnesord og den svenske oversettelsen av DDC[2], CrissCross-prosjektet der tyske emneord er mappet til DDC[3] og FinnONTO (National Semantic Web Ontology Project in Finland)[4].

Det verdensomspennende klassifikasjonssystemet DDC er på god vei til å få sin fullstendige norske oversettelse[5]. Med dette blir mulighetene for å mappe norsk terminologi og emnesystemer til Dewey bedre. I tillegg gir det muligheter for at man i framtiden kan bruke DDC som grunnlag for kobling fra norske emneord til emneord på andre språk som allerede er mappet til Dewey. Flerspråklig søking på tvers av systemer kan bli resultatet for sluttbrukerne.

Etter at vårt prosjekt startet opp, har UBO fått midler til å se på metoder for mapping mellom tesaurusen HUMORD[6] og norsk WebDewey. Samtidig utreder UBO og Nasjonalbiblioteket i et felles forprosjekt muligheten for å etablere en universell norsk tesaurus[7]. Vårt prosjekt vil forhåpentlig kunne bidra med metodikk til mappingen HUMORD-DDC og ser ut til å føye seg inn i rekken av

[1] http://www.ub.uio.no/om/tjenester/emneord/realfagstermer.html
[2] http://www.kb.se/katalogisering/Klassifikation/ddk/Mappning-SAO-Dewey/
[3] http://linux2.fbi.fh-koeln.de/crisscross/index_en.html
[4] http://www.seco.tkk.fi/projects/finnonto/
[5] Norsk WebDewey, http://www.nb.no/Bibliotekutvikling/Kunnskapsorganisering/WebDewey
[6] http://www.BIBSYS.no/files/out/humord/
[7] http://www.nb.no/Bibliotekutvikling/Kunnskapsorganisering/Tesaurus-forprosjekt

prosjekter som peker ut en ny kurs i emne- og klassifikasjonsarbeidet i Norge med ønske om samordning og samvirke.

1.2 Mål

Prosjektet har hatt som hovedmål å utrede metoder for mapping av terminologi i emneordssystemene Realfagstermer og TEKORD mot klasser i DDC.

Et delmål var å undersøke muligheter for mapping mellom UDC[8] og DDC.

Videre skulle prosjektet kvalitetssikre data fra prosjektet *Realfagstermer og TEKORD[1],* samt planlegge og utvikle eventuelt videre samarbeid.

For å spesifisere problemstillingene underveis i prosjektet formulerte prosjektgruppen følgende spørsmål:

- Hva skal mappes, og hvor starter vi når vi har et emnesystem uten struktur, der vi ikke har noe fagområde eller hierarki å ta utgangspunkt i?
- Kan en automatisk kobling av termer i TO/RT mot termer i norsk WebDewey gi oss et godt utgangspunkt for mapping?
- Kan en slik automatisk kobling lette det manuelle arbeidet?
- Hvilke metoder skal vi bruke i det intellektuelle arbeidet med mapping?
- Hvordan kan vi få overganger mellom UDC og DDC?

1.3 Prosjektgruppen

Prosjektgruppen har bestått av følgende representanter:

Ellen M. S. Flatby (universitetsbibliotekar, NTNU UB)

Mari Lundevall (spesialbibliotekar, UBO Realfagsbiblioteket)

Dan Michael O. Heggø (rådgiver, UBO Realfagsbiblioteket)

Heidi Sjursen Konestabo (førstebibliotekar, UBO Realfagsbiblioteket)

Kyrre Traavik Låberg (overingeniør, UBO Realfagsbiblioteket)

Viola Kuldvere (prosjektleder, UBO Realfagsbiblioteket)

Gruppen har hatt felles møter og ellers kommunisert per e-post og Google Docs. To av gruppemedlemmene deltok på EDUG-møtet i Reykjavik 22.-23. mai 2014, og ga en kort presentasjon av prosjektet under møtet i EDUGs mappinggruppe.

[8] http://www.udcc.org/index.php/site/page?view=about

2 Datagrunnlag

2.1 TEKORD

TEKORD er en kontrollert, hierarkisk emneordsliste som er utviklet ved NTNU UB og brukes for de tekniske og naturvitenskapelige fagene. Listen inneholder overordnede og underordnede termer (OT og UT), kvalifikatorer, se-henvisninger og se også-henvisninger. Hvert emneord er koblet til en UDC-klasse.

TEKORD oppdateres etter behov, men tilveksten pr år er liten, fordi det allerede er bygd opp et bredt utvalg av termer. Fagansvarlig innen hvert fagområde lager nye emneord etter behov og finner tilhørende klassifikasjon fra UDC online[9]. Fagansvarlige har kontakt med de enkelte fagmiljø ved NTNU, slik at emneordene blir i overensstemmelse med fagterminologien.

Til prosjektet har vi brukt TEKORD uttrykt i SKOS-modellen[10]. RDF/XML-dumpen vi har brukt ble laget fra data fra BIBSYS emnemodul 4. mars 2012 og inneholder 15374 begreper. Vi har antatt at TEKORD er statisk nok til at det ikke gjør noe at den ikke er helt fersk.

2.2 Realfagstermer

Realfagstermer er et kontrollert vokabular utarbeidet av Realfagsbiblioteket. Vokabularet er gruppert i innholdsbeskrivende emneord, og emneord for form, sted og tid. Alle de fire typene kan, i tillegg til en foretrukken term, ha synonymer, oversettelser og/eller akronymer. I tillegg kan det være registrert klassifikasjon (DDC eller MSC), se også-henvisninger, noter (til internt bruk) og definisjoner (tiltenkt sluttbruker). Hvert begrep, uttrykt ved minimum en foretrukken term, har en unik identifikator.

Enkelttermer kan være kombinert til strenger. Hver streng har også en unik identifikator.

Realfagstermer vedlikeholdes ikke i BIBSYS emnemodul, men i et lokalt system. Navigasjon og indeksering gjøres med husprogramvaren Roald[11]. Til prosjektet har vi brukt Realfagstermer uttrykt i SKOS, som eksportert fra Roald den 8. april 2014. Materialet inneholder 15054 begreper og 15961 strenger.

2.3 TORT

TORT er navnet på overgangen mellom Realfagstermer og TEKORD, altså settet av mappinger mellom de to vokabularene. Overgangen definerer TORT-vokabularet, et virtuelt, felles vokabular, som per i dag består av 3161 begreper fra mappingene etablert i vårt foregående prosjekt[1].

Mappingene ble automatisk generert på bakgrunn av tekstlikhet mellom termer, og det ble tatt stikkprøver for å vurdere kvaliteten. I dette prosjektet har vi fortsatt arbeidet med å kvalitetssjekke mappingene, og begynt arbeidet med å bestemme relasjonstyper. Videre har vi prioritert begreper i TORT når vi har arbeidet med mappinger fra RT til DDC.

2.4 DDC23

[9] http://www.udc-hub.com/
[10] http://www.ntnu.no/ub/data/tekord/
[11] http://folk.uio.no/knuthe/progdist/

Fra Nasjonalbiblioteket har vi fått deler av den norske Dewey-oversettelsen til prosjektformål. Vi begynte med 500-gruppa, og har siden fått 600-640. Det er viktig å understreke at vi har fått oversettelsen med forbehold om at det ikke er den endelige versjonen, og i forståelse om at den ikke publiseres videre. Disse dataene fikk vi som XML-filer uttrykt i MARC21 Authority og MARC21 Classification. Materialet inneholder 3347 klasser fra 500-gruppen og 4166 fra 600-gruppen, men ikke alle er ferdig oversatt. Siden vi hadde begge de to andre vokabularene uttrykt i SKOS laget vi et script[12] som konverterte Deweydataene fra MARC21 Classification til SKOS. MARC21-formatene er langt mer uttrykksfulle enn SKOS, så konverteringen er ikke triviell. Vi gikk for en pragmatisk konvertering der vi fokuserte på å få oversikt over hvilke felt vi trengte, og lagde regler for å konvertere disse (se [12] for detaljer). Uttrykt i SKOS ble filene mindre og kjappere å jobbe med.

2.5 Katalogdata

Fra BIBSYS fikk vi en dump med klassifikasjon og emneord for alle katalogposter fra tidenes morgen (1966) og frem til 28. april 2014, levert som BIBSYS-MARC.

3 Utvikling av metoder

3.1 Mapping

Et viktig verktøy har vært *ISO 25964 Information and documentation : thesauri and interoperability with other vocabularies*. Vi har særlig brukt del 2, som kom ut våren 2013: *Interoperability with other vocabularies*[3, 4].

CrissCross-prosjektet i Tyskland (2006–2010) er det største prosjektet vi kjenner til som har mappet emneord til DDC. Fokuset var på mapping av de kontrollerte emneordene Schlagwortnormdatei (SWD) som i dag er integrert i Gemeinsame Normdatei (GND). For å uttrykke graden av samsvar mellom et emneord og en klassekode, brukte man en skala fra 1 til 4, der 4 betegnet det sterkeste samsvaret. Det ble utarbeidet retningslinjer for mapping [5][13].

Prosjektet var meget omfattende, og brukte ingen automatisk mappemetode. En antydning av omfanget er at til 500-gruppen alene ble det etablert 33438 mappinger fra GND til DDC. Prosjektet er avsluttet, men mappearbeidet pågår fremdeles.

3.1.1 Definisjoner

Definisjonene nedenfor er våre oversettelser fra definisjonene i ISO 25964-2[3], og tallet i parentes refererer til nummereringen der.

Mappe (verb) - å etablere relasjoner mellom begrep i ett vokabular og begrep i et annet vokabular. Alternative uttrykk for dette kan være å koble, eller å etablere overganger. (3.39)

Mapping (verbalsubstantiv) - aktiviteten å mappe. (3.40)

Mapping (substantiv) - er produktet av mappeprosessen: En relasjon mellom et begrep i ett vokabular og begrep i et annet vokabular. Alternative uttrykk kan være relasjoner, eller koblinger. (3.41)

[12] https://github.com/scriptotek/mc2skos
[13] også kort beskrevet på http://www.webcitation.org/6Qjw32FYd (Arkivert fra http://linux2.fbi.fh-koeln.de/crisscross/degrees_of_determinacy_en.html)

Overgang (crosswalk) – et sett av mappinger mellom to eller flere vokabularer. (3.21)

Kildevokabular - vokabularet som tjener som utgangspunkt når man leter etter en korresponderende term eller begrep i et annet vokabular.(3.72)

Målvokabular - vokabularet man slår opp i for å finne korresponderende term eller begrep til term eller begrep fra kildevokabularet. (3.82)

Enveis mapping - mappinger som er etablert i bare én retning, i motsetning til, og mindre arbeidskrevende enn, toveis mapping. (6.3)

En-til-en-mapping - når enkeltbegrep fra ett vokabular mappes til enkeltbegrep i et annet vokabular. Et enkelt begrep kan ha flere mappinger. (Dette er motsatt *en-til-mange-mapping*, der enkeltbegrep i ett vokabular mappes til en kombinasjon av to eller flere begrep i et annet vokabular). (3.55, 3.56)

Mappekandidater - begrepspar som skal vurderes med tanke på å etablere en mapping. (14)

Mapping mellom sammensatte uttrykk (compound equivalence) - mapping mellom begrep der ett begrep i det ene vokabularet er representert ved to eller flere begrep i det andre vokabularet. (3.14)

3.2 Formelle avgrensinger

I henhold til standarden har vi valgt følgende: Kildevokabularet vårt er Realfagstermer, med særskilt vekt på delmengden TORT. Målvokabularet vårt er DDC. Vi har valgt en enveis en-til-en-mapping. Vi har valgt å utelukke mapping mellom sammensatte uttrykk i denne omgang.

Vi har valgt å bruke følgende relasjonstyper, hentet fra ISO 25964-2 (tallet i parentes refererer til kapitlet der relasjonstypen beskrives i standarden):

- **=EQ** Eksakt samsvar: En samsvarsrelasjon mellom to begrep innebærer at de to begrepene kan bytte plass med hverandre, uten at det går ut over betydningen. I denne rapporten skriver vi normalt bare EQ (likhetstegnet er altså underforstått). (11.2)
- **~EQ** Tilnærmet eksakt samsvar. Typiske tilfeller: Begrepene kan samsvare i noen tilfeller, men ikke i andre. Begrepene kan ha overlappende betydning, eller små forskjeller i konnotasjon. Når man mapper mellom et klassifikasjonsskjema og en tesaurus er det vanlig å finne en klasse der klassebetegnelsen er lik en foretrukken term, men der nærmere undersøkelser viser at de to ikke er eksakt like. (11.3)
- **BM** Hierarkisk relasjon der kildebegrepet er smalere enn målbegrepet. (9)
- **NM** Hierarkisk relasjon der kildebegrepet er videre enn målbegrepet. (9)
- **RM** Assosiativ relasjon: Kildebegrepet kan assosieres med målbegrepet på en slik måte at man kan anta at målbegrepet vil være relevant for den som søker etter kildebegrepet. (10)

3.3 Automatiske mappeforslag

Kildevokabularet vårt, Realfagstermer, er stort, og siden det ikke er noen tesaurus har vi ikke en faglig oppdeling som kunne være egnet for å finne startpunkter til prosjektet. Derfor var det en tidlig

utfordring rett og slett å finne ut hvor vi skulle begynne. Det lå fast før vi startet at vi ønsket å finne semiautomatiske metoder, altså var et rent manuelt prosjekt utelukket. I stedet har vi valgt å prøve to ulike, maskinassisterte tilnærminger for å plukke ut kandidater til mapping: Den ene er en tekstlig sammenlikning mellom Realfagstermer og DDC, den andre er forslag fra katalogdata, basert på samtidige forekomster av emneord og klassenummer.

3.3.1 Tekstsammenlikning

Den tekstlige sammenlikningen er begrenset til eksakt strengmatching. Under overskriftene Test 1, 2 og 3 under har vi beskrevet den iterative prosessen vi har vært gjennom for å komme frem til Test 4, som har generert resultatene vi har brukt videre i prosjektet. Prosessen har vært preget av at vi har fått tilgang til mer DDC-data og at vi har fått en bedre forståelse av MARC21 Classification-modellen. Gjennom å bruke RDF-utgaven av Realfagstermer har vi også funnet småfeil i RDF-eksporten som vi har fått rettet.

Test 1

I første forsøk sammenliknet vi RT-begrep med registertermer i DDC. Resultatet var 1594 treff, og for alle treff ble det skrevet ut en del opplysninger vi hadde spesifisert.

En gjennomgang av testen avdekket noen problemer med konverteringen fra MARC21 Classification til SKOS, og med spørringene. Vi hadde bare brukt foretrukne termer fra RT, bare fått med første registerterm for hvert klassenummer, fikk tilsynelatende uforklarlige treff som viste seg å være bygde nummer, og det var en utfordring å bruke strengene våre som utgangspunkt.

Vi avgjorde at vi ville gå direkte videre med en ny test, uten å bruke resultatet til videre mapping.

Test 2

Med en ervervet bedre kjennskap til dataformatet i Deweymaterialet, gjorde vi neste test: RT-begrep, både foretrukne og ikke-foretrukne termer, ble sammenliknet mot klassebetegnelser og registertermer. Vi forsøkte også å inkludere strenger i søket, men lyktes ikke med å få noe treff i strenger. Resultat: 2470 RT-begrep traff 1766 DDC-nummer. Vi rangerte ikke mappingkandidatene etter hvorvidt treffene fantes fra foretrukne eller ikke-foretrukne termer.

Fremdeles var imidlertid resultatet noe uoversiktlig å jobbe videre med.

Test 3

Vi forenklet testen, og sammenliknet RT, både foretrukne og ikke-foretrukne termer, med klassebetegnelser i DDC. Vi fikk generert 972 mappingforslag fra 804 RT-begrep til 970 DDC-klasser.

Test 4

Dette var en gjentagelse av test 3, men med oppdaterte RT-data, og inkludert materiale fra 600-gruppa. Vi fikk nå 1681 mappingforslag fra 1200 RT-begrep til 1674 DDC-klasser. Resultatet ble delt inn i lister som vi har behandlet videre i et verktøy som er nærmere beskrevet i kapittel 4.

3.3.2 Statistisk sammenlikning

Med utgangspunkt i katalogdata fra BIBSYS gjorde vi et utplukk for å identifisere mappekandidater. Vi valgte å begrense oss til katalogposter for dokumenter klassifisert ved UBO i perioden januar 2012 til april 2014. Den relativt korte perioden er begrunnet med at dette er perioden vi har brukt Realfagstermer som kontrollert vokabular, og vi forventer derfor at emneordspraksisen har høyere kvalitet i denne perioden enn tidligere.

Inkludert i analysen var katalogposter med minst ett UBO-RT-emne (687k $2 no-ubo-mn) og minst én UBO-DDC23-klassifikasjon (082k $2 DDC-23). Det ble funnet 1639 slike poster katalogisert i utvalgsperioden. Disse postene hadde 1816 unike RT-emneord brukt til sammen 3307 ganger, og 934 unike DDC-klasser brukt til 1864 ganger. Hver post har altså i snitt rundt én DDC-klasse og to RT-emneord. Tabell 3.1 og 3.2 viser hvor mange ganger emneordene og klassene i utvalget vårt er brukt.

Tabell 3.1 Frekvensfordeling for RT-emneord i utvalget

Bruksfrekvens	1	2	3-9	10+	Til sammen
Antall emneord	1222	308	267	19	1816

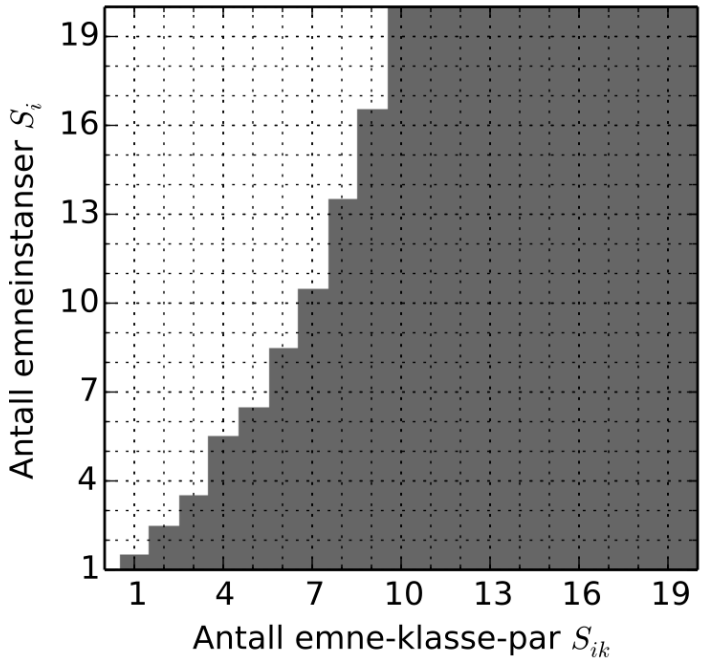
Tabell 3.2 Frekvensfordeling for DDC-klasser i utvalget

Bruksfrekvens	1	2	3-9	10+	Til sammen
Antall klasser	625	143	149	17	934

Vi regnet ut korrelasjonen K mellom et emneord i og en klassifikasjon k som antall ganger emneordet og klassifikasjonen er brukt sammen, S_{ik}, delt på antall ganger emneordet er brukt totalt, S_i ;

$$K = S_{ik} / S_i$$

Som vi ser fra tabell 3.2 er en overveiende del av emneordene i utvalget bare brukt én gang. For å få en viss mengde mappingforslag har vi valgt å prøve å inkludere disse til tross for at vi ikke kan snakke om korrelasjon basert på én katalogpost. Mer spesifikt har vi generert mappingforslag hvis $K \geq 1/(1+0.05S_i)$. Faktoren er valgt basert på litt prøving og feiling, og tanken har vært at kravet til K skal synke med antall dokumenter. Hvis et emneord er brukt på 20 dokumenter, holder det at 10 har brukt en gitt klassifikasjon for å gi assosiasjon, se figur 3.1. Se også diskusjonen i avsnitt 4.5.3.



Figur 3.1 Illustrasjon av $K \geq 1/(1+0.05S_i)$. Punkter i det grå feltet gir mappingforslag.

3.4 Intellektuelle vurderinger

Vi har behandlet mappemetodene og stikkprøver av tilhørende resultater manuelt, etter å ha diskutert mulige valg av relasjoner. Vi har valgt følgende gang: Vi vurderer forslag til relasjon ved hjelp av eventuell ekstrainformasjon om emnebegrepet (henvisninger, engelsk språk, hierarkisk plassering i TEKORD for TORT-ord), og ved hjelp av oppslag i DDC. Videre har vi slått opp i katalogen for å sjekke bruken av både emneord og klassifikasjon. Vi velger så type relasjon. Ideelt skal minst to personer vurdere samme forslag til relasjon. Dersom flere enn en person velger samme type relasjon, er det en godkjent mapping. Dersom det imidlertid velges en annen type relasjon, vurderes mappingen på nytt.

Underveis i prosjektet gjorde vi et valg om å prøve ut følgende hypotese: Dersom vi med metoden Tekstsammenlikning bare får én kobling fra Realfagstermer til DDC (kildebegrepet har kun én foreslått kobling til DDC), er sjansen for at relasjonen er EQ, relativt høy. Dette gjelder for listene A og B i avsnitt 4.1.

Vi har derfor valgt å behandle RT med ett treff og RT med flere treff mot DDC ulikt. Ulikheten består i at vi har satt relasjonen mellom RT og DDC, for alle RT med bare ett treff mot DDC, lik EQ. Deretter har vi gjort en individuell vurdering ved stikkprøver. For alle RT med mer enn én kobling mot DDC, skal alle relasjoner vurderes intellektuelt.

En foreløpig arbeidsbeskrivelse finnes bakerst i denne rapporten.

4 Resultater og diskusjon

4.1 Forslag til mapping

Vi har totalt 2804 mappekandidater, det vil si maskinelt foreslåtte mappinger der RT er kildevokabular og DDC er målvokabular. Et begrep kan være mappet til flere klasser.

Fra metoden Tekstsammenlikning, test 4 (beskrevet i kapittel 3.3.1) fikk vi 1681 forslag til RT-DDC-mappinger basert på eksakt ordlikhet mellom RT-term (prefLabel eller altLabel) og 153 $j (klassebetegnelse) i WebDewey-data. 673 av RT-begrepene som forekommer her, tilhører fellesvokabularet TORT. De 1681 mappekandidatene fordeler seg på 972 forslag fra 500-gruppen og 709 forslag fra 600-640.

Fra metoden Statistisk sammenlikning (beskrevet i kapittel 3.3.2), har vi 1169 forslag til RT-DDC-mappinger basert på korrelasjon i katalogposter for dokumenter klassifisert ved UBO i perioden 2012-01 til 2014-04.

Mappekandidatene ble delt opp i lister avhengig av hvilke parametere de var basert på og hvilken metode de var fremkommet under.

Liste A: Kildebegrepet har kun én foreslått mapping mot DDC
Listen består av 939 RT-DDC-mappinger der RT-begrepet ikke har andre mappingforslag mot DDC. Disse har fått foreslått status EQ. Manuelle (intellektuelle) stikkprøver viser følgende: 118 EQ, 3 ~EQ, 1 BM, 3 NM, 2 avslått.

Liste B: Kildebegrepet har kun én foreslått mapping mot DDC, og er en del av TORT
Dette er en delmengde av liste A, som begrenser seg til mappingforslag der kildebegrepet (i RT) også har «exact match» til begreper i TEKORD. Siden termer i TEKORD er koblet mot UDC, har vi her også en indirekte kobling mellom UDC og DDC.

Her har vi 510 maskinelt foreslåtte mappinger med status EQ. Det er tatt manuelle stikkprøver som viser 110 EQ, 3 ~EQ, 1 BM, 3 NM, 2 avslått. Koblingene mellom UDC og DDC er foreløpig ikke intellektuelt kontrollert.

Eksempler fra liste A og B:

RT *Lava* EQ DDC23: 552.22 *Lava*

RT *Ordinære differensialligninger* EQ DDC23: 515.352 *Ordinære differensialligninger*

RT *Oljeforurensning* ~EQ DDC23: 628.16833 *Oljeutslipp*
Kommentar: RT *Oljeforurensning* har *Oljeutslipp* som synonym

RT *Elektrisitet* NM DDC23: 622.48 *Elektrisitet*
Kommentar: Her er klassenummeret i DDC knyttet til gruvedrift, mens RT *Elektrisitet* har videre betydning.

RT *Atlanterhavslaks* BM DDC23. 597.56 *Laks*

RT *Linser* AVSLÅTT DDC23: 635.658 *Linser*
Kommentar: Disse er homonymer med ulik betydning.

Liste C: Kildebegrepet har mer enn én foreslått mapping mot DDC
Her er det 742 foreslåtte relasjoner. 70 er vurdert så langt, og fordeler seg slik: 22 EQ, 10 ~EQ, 28 NM, 10 avslått.

Liste D: Kildebegrepet har mer enn én foreslått mapping mot DDC, og er en del av TORT
Dette er en delmengde av liste C. RT-begrepene i denne lista er foreslått mappet mot flere enn en DDC-klasse, og er samtidig mappet mot TEKORD. Her har vi derfor også fått en indirekte kobling mellom UDC og DDC.

Her har vi 477 foreslåtte relasjoner. 66 relasjoner er vurdert så langt, og fordeler seg slik: 20 EQ, 8 ~EQ, 0 BM, 28 NM, 10 avslått. Koblingene mellom UDC og DDC er foreløpig ikke intellektuelt kontrollert.

Eksempler fra liste C og D:

RT *Antenner* foreslås mappet mot henholdsvis DDC 621.384135 *Antenner* og 621.38835 *Antenner*. Emneordet har ingen henvisninger eller andre forklarende elementer, det er kun ved å se på bruken at vi kan fastslå betydningen det har i RT. Vi ser da at det er brukt i en teknologisk kontekst, ikke biologisk som kunne vært et alternativ. De to DDC-numrene henviser til henholdsvis radioantenner og fjernsynsantenner. Begge numrene faller da innenfor betydningen av emnebegrepet *Antenner*, men ingen dekker det helt. Vi velger relasjonen NM i begge tilfeller.

RT *Terapi* foreslås mappet mot følgende DDC: 616.891, 616.69206, 616.8906, 618.17806, 618.9706. Emneordet har to ikke-foretrukne termer: *Behandling, Medisinske behandlinger.* DDC-numrene er alle knyttet til behandling for gitte tilstander. Valget står da mellom å avvise mapping, eller å sette relasjonen NM, i hvert enkelt tilfelle. Her velger den første som behandler mappekandidatene, å avvise relasjon i alle tilfellene. Dette er en avveining som kan diskuteres: Relasjonen NM kunne vært utnyttet i søk. Her er det imidlertid også tatt hensyn til hvordan emneordet brukes: Dette er et emneord som vi anser helst bør brukes til å avgrense et annet emneord, slik at vi kan finne det som andre ledd i en emnestreng. En kobling av emnestrengen *Infertilitet : Behandling* og 616.69206 kunne vært nyttig å etablere, mens en kobling av emneordet *Behandling* og 616.69206 kan anses å være støy snarere enn til nytte.

RT *Bein* foreslås mappet mot følgende DDC: 573.76, 599.947, 611.71 og 617.471. Emneordet har to ikke-foretrukne termer: *Knokler, Beinvev*. *573.76 Knokler* er underordnet *573 Bestemte fysiologiske systemer hos dyr, regional histologi og fysiologi hos dyr*. Her har vi valgt relasjonen ~EQ. Argumentet vårt for at dette ikke er en EQ, er at knoklene i DDC er avgrenset til dyreknokler, mens emneordet inkluderer menneskebein. Tilsvarende er *611.71 Knokler* underordnet *611 Menneskets anatomi, cytologi, histologi*, altså avgrenset til menneskets knokler. Her har vi også valgt ~EQ.

599.947 Knokler er underordnet *599.94 Antropometri*. Den første som har behandlet disse mappekandidatene har valgt relasjonen NM, men det kan også argumenteres for at relasjonen avvises.

617.471 Knokler er underordnet *617.4 Kirurgi inndelt etter systemer*. Den første som har behandlet disse mappekandidatene har valgt relasjonen NM, men det kan også argumenteres for at relasjonen avvises.

Kan alt mappes? Det ser ut som innholdsbeskrivende emneord av allmenn karakter [6] bare med forsiktighet bør mappes. 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 (f.eks. *Fugler, Dobbeltbindinger, Trådløs kommunikasjon*) og innholdsbeskrivende emneord av allmenn karakter (f.eks. *Metoder, Utvikling, Vekst*). Dette er en mangel ved vokabularet som kan forklares ved hvordan det er laget, og som det vil ta tid å rydde opp i. Sannsynligvis er det lurt å veie fordeler og ulemper sterkt mot hverandre før man mapper de mest allmenne emneordene.

Liste E: Mappekandidater basert på statistisk sammenlikning

Listen består av 1169 forslag til RT-DDC-mappinger basert på metoden statistisk sammenlikning (som beskrevet i 3.3.2) i katalogposter for dokumenter klassifisert ved UBO i perioden 2012-01 til 2014-04. Her må alle relasjoner gjennomgås manuelt. Vi har så vidt startet på dette arbeidet og viser her kun noen eksempler.

Eksempler på foreslåtte relasjoner (relasjoner som enda ikke er intellektuelt vurdert):

RT *Statistikk* FORESLÅTT DDC23: 519.5 *Matematisk statistikk*
Kommentar: Deweynummeret er brukt på 9 av 15 dokumenter med denne Realfagstermen.

RT *Akvariefisker* FORESLÅTT DDC23: 639.34 *Fiskeoppdrett i akvarier*
Kommentar: Deweynummeret er brukt på 1 av 1 dokumenter med denne Realfagstermen.

RT *Katter : Kjæledyr* FORESLÅTT DDC23: 636.8 *Katter*
Kommentar: Deweynummeret er brukt på 2 av 2 dokumenter med denne Realfagstermen.

RT *Solenergi*
RT *Energiressurser*
RT *Miljøvern*
RT *Vindkraft*
RT *Kjernekraft*
RT *Vannkraft*
RT *Bølgekraft*
Alle FORESLÅTT til DDC23: 531.6 *Energi*

Eksempler på intellektuelt vurderte relasjoner:

RT *Solceller* BM DDC23: 621.381542 *Fotoelektriske og fotoelektroniske komponenter*. Deweynummeret er brukt på 1 av 1 dokumenter med denne Realfagstermen. *Solceller (photovoltaic cells)* finner vi i en inkluderer-note i WebDewey. Betydningsinnholdet i RT er altså smalere enn i Deweyklassen. Derfor settes relasjonen ikke til EQ, men til BM.

RT *Kvanteteori* EQ DDC23: 530.12 *Kvantemekanikk (kvanteteori)*. Deweynummeret er brukt på 6 av 8 dokumenter med denne Realfagstermen.

RT *Bartrær* ~EQ DDC23: 585 *Pinophyta (nakenfrøete planter):* Deweynummeret er brukt på 1 av 1 dokumenter med denne Realfagstermen. Det er en her-note for termen i DDC. I følge CrissCross-prosjektets retningslinjer[5] kan ikke en her-note være EQ (D4), da relasjonen skal kunne gjelde begge veier og begrepene skal kunne byttes ut med hverandre. Den kan ofte være D3, tilsvarende ~EQ. Det kan også diskuteres om relasjonen kan settes til BM. Vi har først også vært inne på EQ fordi en her-note i DDC indikerer at termen er ekvivalent til hele klassen, og sidestiller dermed termene.

4.2 Verktøy

4.2.1 Bakgrunn: Regneark

I begynnelsen arbeidet vi med mappingtabeller i regneark, med kolonner for kildebegrep, målbegrep, relasjonstype, kommentar og signatur, og én rad per mapping. Én person behandlet hver rad og signerte med sine initialer.

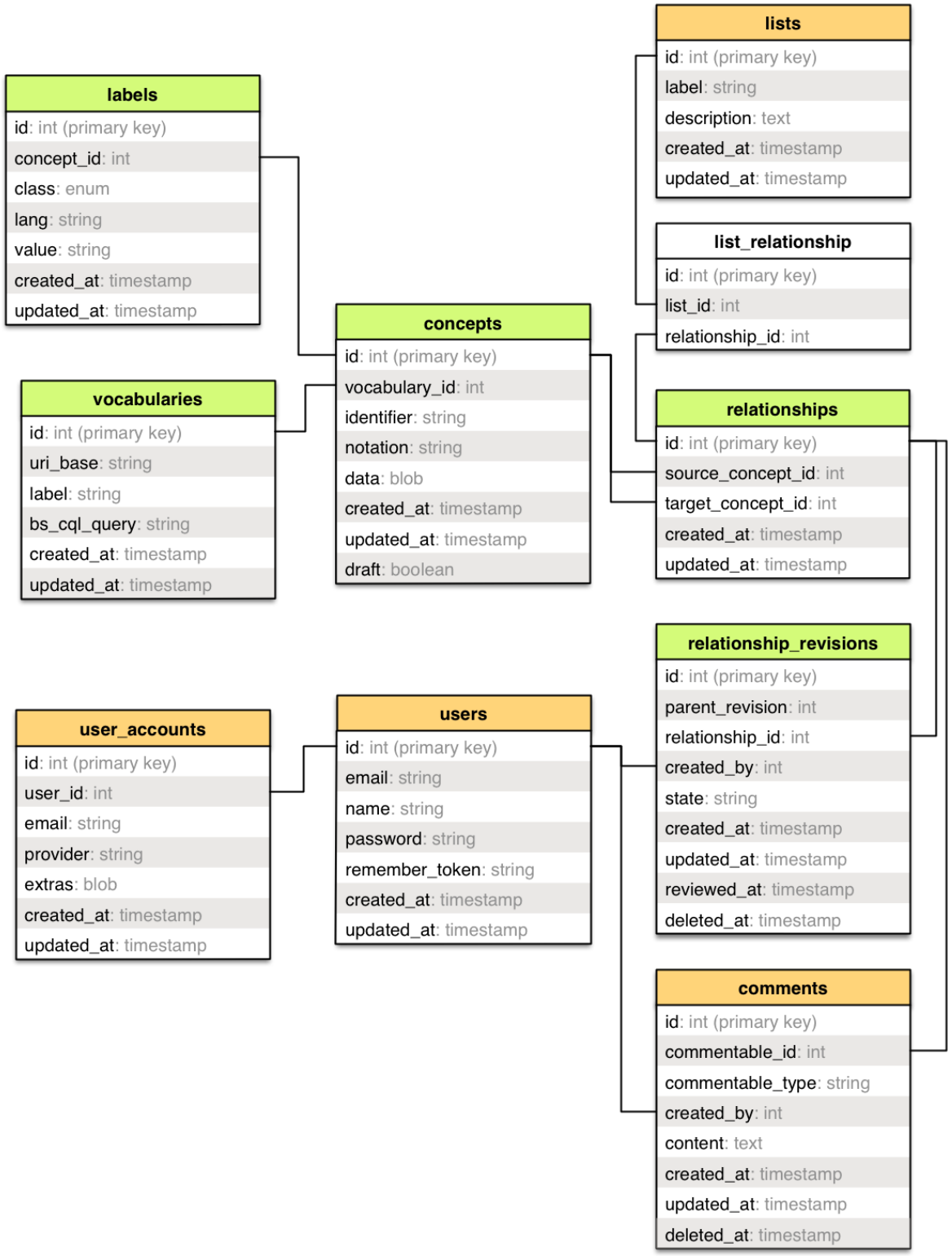
Det flate tabellformatet skapte noen uønskede begrensninger. Ved behandling av en bestemt mapping, så vi for eksempel at det ville vært fordelaktig enkelt å kunne se alle mappingene til de involverte begrepene, ikke bare den éne man behandler. Hvis flere personer skulle kommentere den samme mappingen ble det også fort uoversiktlig. Videre savnet vi en oversiktlig endringshistorikk, og vi hadde en viss bekymring for dataintegriteten. Uten validering av dataene kunne en uten problemer definere en mapping mellom to ikke-eksisterende begreper, og dessuten er det litt for lett å komme borti en tast i feil celle og overskrive en verdi i et regneark uten at en legger merke til det.

4.2.2 µmapper: Datamodellen

Et stykke ut i prosjektet bestemte vi oss derfor for å forsøke å utvikle et verktøy, µmapper[14], for å jobbe med mappingene. Som datalager valgte vi en MariaDB-database med InnoDB-tabeller, fordi det var et system vi kjente godt, og fordi det gir god støtte for dataintegritet direkte i databaselaget. Tabellstrukturen vår er vist i figur 4.1.

En innførsel i tabellen *relationships* definerer at en relasjon (eller mapping) finnes mellom to begreper i ulike vokabularer, men ikke hvilken type relasjon. Mens selve relasjonene i dette prosjektet i hovedsak har blitt opprettet maskinelt, har relasjonstypene vært noe vi har satt manuelt. Siden dette er et intellektuelt krevende arbeid, er det interessant å bevare mest mulig historikk knyttet til prosessen, noe vi har gjort gjennom kommentarer og relasjonsrevisjoner. Hver relasjon har minst én relasjonsrevisjon (*relationship_revisions*), som koder informasjon om relasjonstype, hvem som har behandlet den og et tidsstempel. Hver gang relasjonstypen endres, opprettes en ny revisjon. I grensesnittet vises informasjonen som en logg, se figur 4.2.

[14] https://github.com/scriptotek/mumapper



Figur 4.1 Tabellstruktur for µmapper. Tabeller som brukes for å lagre kunnskapsorganisasjonsdata har oransje overskrift, mens metatabeller som bare brukes for å støtte intern arbeidsflyt har grønn overskrift. Tabeller for mellomlagring og statistikk er utelatt. Kommentarer kan i prinsippet hektes på flere modeller, men har i praksis bare blitt brukt til relasjoner.

Figur 4.2 Historikk for én relasjon med tre revisjoner og to kommentarer i µmapper.

Settet av relasjonstyper er forhåndsdefinert og består av typene anbefalt i ISO-en (eksakt samsvar, nær-eksakt, bredere, smalere og relatert), samt arbeidstypene «foreslått» og «avslått». Relasjonstypen «foreslått» settes som standard for relasjoner opprettet gjennom automatisk mapping. Når slike automatiske mappingforslag intellektuelt blir vurdert som feilaktige, brukes relasjonstypen «avslått». Dette forklarer automatiske mappeprosesser at relasjonen ikke skal forsøkes opprettet på nytt i fremtiden. De to relasjonstypene «foreslått» og «avslått» er altså interne arbeidsverktøy, og slike relasjoner inkluderes ikke i dataeksporter.

Begrepene er definert ved deres URI-er, men for å gjøre dem gjenkjennelige for mennesker har vi hentet inn termer, klassebetegnelser og overordnete begreper fra begrepenes autoritetsdata, som mellomlagres i databasen vår for rask visning. Denne informasjonen må kunne oppdateres med nye autoritetsdata ved behov.

4.2.3 µmapper: Applikasjonen

For å gjøre utviklingstiden så kort og forutsigbar som mulig, brukte vi et rammeverk vi kjente fra før, PHP-rammeverket Laravel. Fordelaktig funksjonalitet i rammeverket inkluderer den objektrelasjonelle mapperen Eloquent, som abstraherer vekk databasekallene, og et system for migreringer, så en får versjonhistorikk på databaseskjemaet. På toppen av dette utviklet vi en applikasjon for å opprette, redigere og kommentere relasjoner/mappinger.

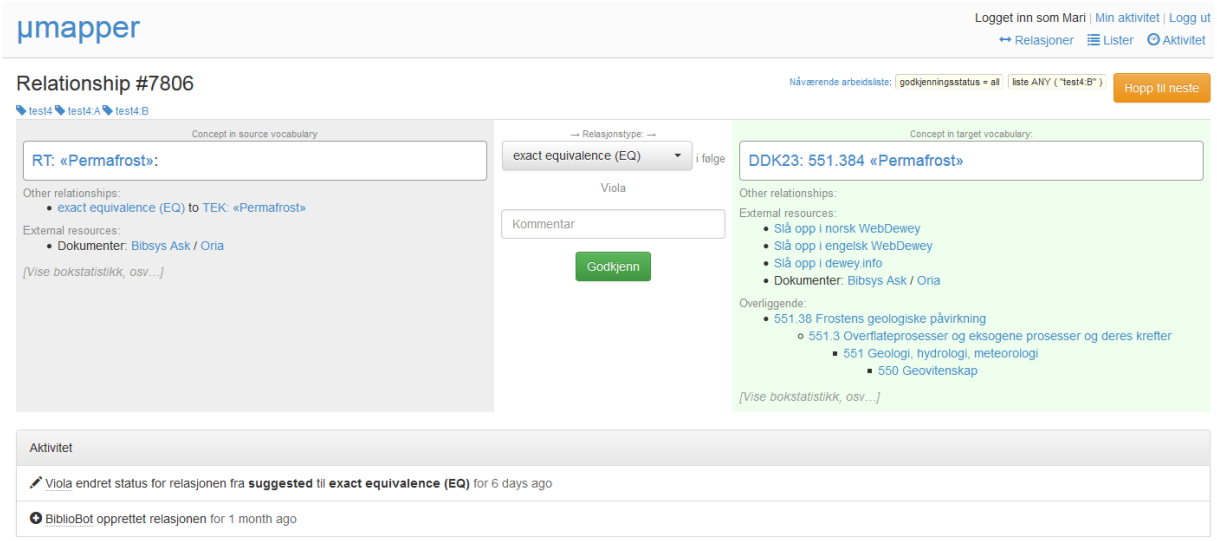
µmapper er avgrenset til å være en ren mappingbase, det er ikke et verktøy for å faktisk generere mappingforslag. Mappingforslag finnes av frittstående script som lagrer dem i µmapper gjennom et JSON-API, og µmapper forholder seg til et mappingforslag lagt inn av et script på samme måte som et forslag lagt inn av et menneske. Med frittstående script er det enkelt å utvikle og teste nye script uten å forholde seg til hvordan µmapper til enhver tid virker.

I dette prosjektet har vi utviklet to enkle script for å finne mappingforslag basert på metodene i 3.1.1 og 3.3.2, samt et script som har importert mappingforslag fra TORT.

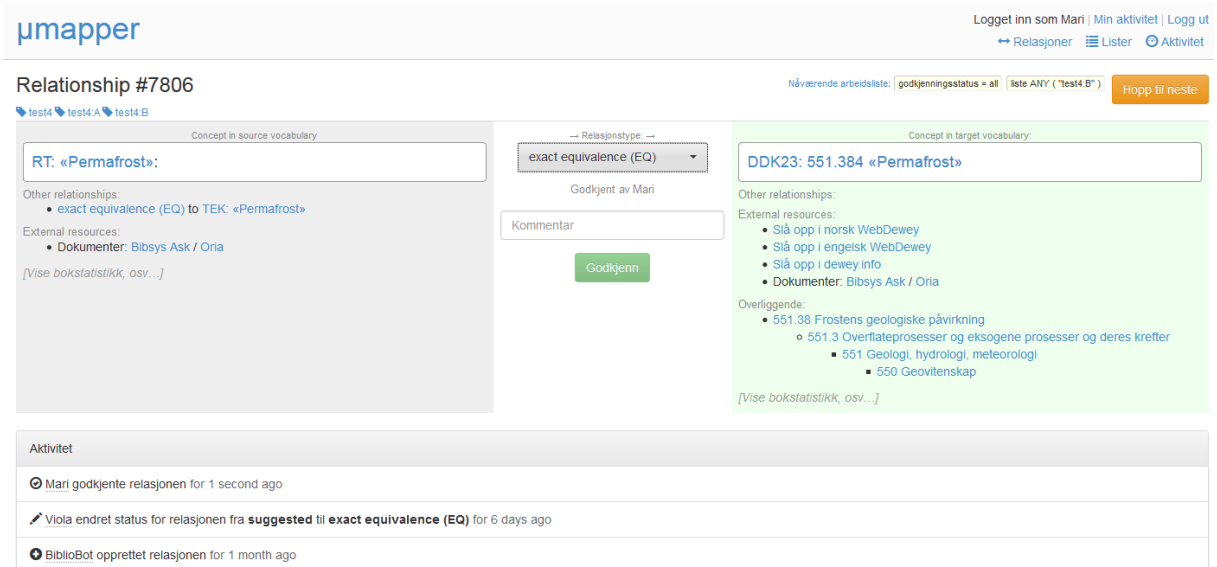
4.3 Arbeidsbeskrivelse for mapping

Figur 4.3-4.4 viser eksempler på hvordan vi konkret har jobbet med et mappingforslag i µmapper, mens figur 4.5-4.6 viser den endelige RDF-representasjonen som vi anser som et sluttprodukt. Vi kommer til å publisere hele overgangen som åpne, lenkede data på Realfagtermers nettsted.[15]

For å legge til rette for en mest mulig enhetlig vurdering av relasjonstyper har vi laget et utkast til arbeidsbeskrivelse (se side 23). Praksisen og arbeidsbeskrivelsen vil også bli tatt opp til diskusjon og videreutvikling i UBO-prosjektet *På vei mot en generell norsk tesaurus* hvor metoder for mapping av HUMORD til DDC skal utredes.



Figur 4.3 En mappekandidat, identifisert ved den unike id #7806, mellom kildebegrepet *Permafrost* i Realfagstermer og Deweyklassen 551.384 i DDC 23. Mappekandidaten er automatisk foreslått av et script (BiblioBot). Relasjonstypen er deretter intellektuelt vurdert og valgt fra nedtrekksmenyen i midten av bildet. Under overskriften *RT: «Permafrost»* er det også listet opp andre mappinger fra begrepet. Her ser vi en relasjon til TEKORD.



[15] http://www.ub.uio.no/om/tjenester/emneord/realfagstermer.html

Figur 4.4 Relasjonstypen er intellektuelt vurdert for andre gang, og den valgte relasjonstypen godkjent.

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:skos="http://www.w3.org/2004/02/skos/core#">

  <skos:Concept rdf:about="http://folk.uio.no/knuthe/emne/data/xml/#REAL009111">
    <skos:exactMatch rdf:resource="http://dewey.info/class/551.384/e23/"/>
    <skos:exactMatch rdf:resource="http://ntnu.no/ub/data/tekord#NTUB08394"/>
    <skos:prefLabel xml:lang="nb">Permafrost</skos:prefLabel>
  </skos:Concept>

</rdf:RDF>
```

Figur 4.5 Den samme relasjonen vist i µmapper som RDF/XML. Legg merke til at det ikke står noe om relasjonens status, eller informasjon om historikken/tidsstempling. Vi har foreløpig ikke funnet ut hvordan vi skal uttrykke dette, men ønsker å gjøre det i fremtiden.

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .

<http://folk.uio.no/knuthe/emne/data/xml/#REAL009111>
  a skos:Concept ;
  skos:exactMatch <http://dewey.info/class/551.384/e23/>, <http://ntnu.no/ub/data/tekord#NTUB08394> ;
  skos:prefLabel "Permafrost"@nb .
```

Figur 4.6 Den samme relasjonen vist i µmapper som RDF/Turtle.

4.4 Realfagstermer og TEKORD

4.4.1 Kvalitetssikring av eksisterende data

Det inngikk i prosjektet å kvalitetssikre dataene fra overlappen mellom TEKORD og Realfagstermer, TORT, som består av 3161 begreper som de to emnesystemene har felles. Vi har reetablert SKOS-filen med begrepene fordi Realfagstermer grunnet en større revisjon fikk nye identifikatorer. TORT er gjenopprettet med de nye identifikatorene og blir publisert som lenkede data (RDF/SKOS). Disse dataene kan vi - og andre - nå jobbe videre med for å teste ut eventuell nytteverdi i søk.

I dette prosjektet har vi gått videre med intellektuell vurdering av TORT. Det er gjort på følgende måte: Begrepenes betydningsinnhold er først forsøkt fastslått ved hjelp av henvisninger, og for TEKORDs del dessuten hierarki og UDC-nummer. Deretter er det sett på bruken av emneordene. Relasjonene er foreløpig bare vurdert av én person, og vi har prioritert de begrepsparene der RT-termen samtidig er mappekandidat til DDC. Det gjelder 673 mappinger, der vurderingene er fordelt slik: 634 EQ, 2 ~EQ, 1 BM, 1 NM, 4 avslått. 31 har kommentarer og er ikke ferdig vurdert.

Disse 673 TORT-begrepene er maskinelt mappet mot totalt 987 DDC-nummer (av disse er 283 relasjoner manuelt vurdert), og gir dermed en indirekte kobling mellom UDC og DDC.

4.4.2 Videre samarbeid

Vi har tidligere konkludert med at de to vokabularene ikke slås sammen fordi hovedmengden av termene tilhører ulike fagområder og det vil kreve en innsats som ikke vil stå i forhold til resultatet. Vokabularene nærmer seg allikevel hverandre terminologisk og blir gradvis mer samordnet, siden vi nå har et mer aktivt forhold til hverandres vokabularer. Ved oppretting av nye termer i ett vokabular, er det bestemt at vi skal konsultere det andre vokabularets eventuelle bruk av samme term. Vi har fortsatt stor nytte av forrige prosjekts resultater med henhold til retting og rydding i eget vokabular.

NTNU UB ser spesielt positivt på en eventuell videre mapping UDC-DDC fordi det kan gi bedre innganger til klassifikasjon og emner på e-ressurser, som ofte leveres med Deweydata.

For videre samarbeid vil også utfallet av forprosjektet om en universell norsk tesaurus være av betydning.

4.5 Generell diskusjon

Generelt: Dewey-datagrunnlaget har vært begrenset til en delvis ferdig oversettelse av 500-gruppen og 600-640. Det er klart at betydelige deler av Realfagstermer vil kunne mappes til Dewey-klasser utenfor disse områdene.

4.5.1 Sammenlikning av tekststrenger – metoden

Sammenlikning av tekststrenger har gitt oss mange forslag til mappinger og et godt utgangspunkt for å starte mappingarbeidet. Samtidig ser vi mange veier videre for å utvikle metoden og generere flere mappingforslag.

I dette prosjektet har vi begrenset oss til eksakt strenglikhet. En naturlig fortsettelse vil være å tillate en liten grad av ulikhet gjennom for eksempel et mål for redigeringsdistanse[16].

Videre vil det helt klart være interessant å trekke inn mer data i tekstsammenlikningen. Dewey-materialet er rikt. Vi så for eksempel i test 2 (se 3.3.1) at vi fikk mange treff mot registertermene. Å inkorporere dem i den daværende regneark-arbeidsmåten ble for uoversiktlig, men med mappingverktøyet har vi mulighet til å dele opp trefflisten i logiske lister, samtidig som vi kan få hele bildet når vi trenger det. Videre forventer vi at notene i Dewey-materialet vil være en god kilde til mappingforslag, spesielt «Her»- og «Inkluderer»-notene.

Strenger, både emnestrenger og generelt termer som består av flere ord, representerer en ekstra utfordring. I videre arbeid vil det antakelig være gunstig å «tokenisere» dem; bryte dem opp i ord, og sammenlikne ord for ord – og kanskje akseptere at ikke alle ord er like. 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.

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. Det kan være aktuelt å se på metoder for å redusere mengden støy, f.eks. ved å gjøre en frekvensanalyse på materialet og unngå vanlige ord slik som «metoder».

[16] https://en.wikipedia.org/wiki/Edit_distance

4.5.2 Hypotese

Som beskrevet i kapittel 3, valgte vi å sette relasjonen automatisk til EQ, der det kun var én mappekandidat fra et gitt RT-begrep. Tallene for stikkprøvene viser så langt at vi ikke må avvise hypotesen. Før vi kan konkludere om kvaliteten er god nok må vi gjøre flere stikkprøver, og finne et kvantitativt mål for presisjon.

En svakhet ved resultatene våre er at vi kun har mappet mot deler av DDC (500-gruppen og 600-640). 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. Hvis dette gjelder for en betydelig del av resultatene, vil hypotesen måtte revurderes. 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. For å dekke informatikk må ihvertfall 000-gruppen med.

4.5.3 Mapping basert på statistisk sammenlikning

Som vi beskrev i avsnitt 3.3.2 gjorte vi en avgrensning som medførte at vi fikk et utvalg på kun 1639 katalogposter, og der en overveiende grad av emneord og klasser kun var brukt én gang. Datamaterialet var derfor egentlig for lite til å kunne trekke statistiske slutninger fra det. Vi valgte et enkelt krav til assosiasjon som kunne oppfylles av data fra én enkelt katalogpost. Vi ser at dette har gitt gode forslag, men også støy, og dessverre har vi ikke fått sjekket nok stikkprøver til å vurdere om presisjonsnivået er tilstrekkelig høyt til å kunne forsvare det.

I fremtidig arbeid må vi helt klart se på mulighetene for å utvide datamaterialet, for eksempel gjennom å inkludere et større årsrom, og å inkludere DDC-klassifikasjon gjort av andre enn Universitetsbiblioteket i Oslo. Vi må også utvikle metoden videre, siden den foreløpig er nokså improvisert. Spesifikt bør vi finne et anerkjent statistisk mål på assosiasjon vi kan bruke. En kandidat er Log-Likelihood-koeffisenten, som Library of Congress har brukt i sitt arbeid med å mappe LCSH til DDC, fordi den skal fungere godt med glisne data (*sparse data*) og små datasett [7]. Datasettet deres (~700 000 LCSH-DDC-par) er imidlertid to størrelsesordener større enn vårt (~4000 par), så vi må kontrollere hvordan det fungerer med data i vår størrelsesorden, og eventuelt undersøke andre mål.

4.5.4 Vurdering av relasjonstyper

Vi er fremdeles i diskusjon om hva som er det rette valget av relasjon i noen typer tilfeller.

Å bestemme relasjonstype kan være vanskelig. Vi har hatt gode diskusjoner underveis, uten å ha kommet dithen at vi er enige i alle tilfeller, som eksemplene i avsnitt 4.1 tydelig viser. På et mappingmøte under EDUG-møtet i Reykjavik 23. mai, 2014, fikk vi tatt diskusjonen med flere innen miljøet, og fikk bekreftet hvor vanskelig dette er.

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?

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.

4.5.5 Overganger mellom UDC og DDC

673 TORT-begrep er maskinelt mappet mot totalt 987 DDC-klasser. Siden alle TEKORD er forsynt med UDC, har vi oppnådd en indirekte kobling mellom UDC og DDC. En manuell sjekk av disse koblingene vil være nødvendig og kan utføres ved å behandle mappekandidatene i vårt mappeverktøy. Neste skritt kan være å bruke katalogdataene fra BIBSYS for å få forslag til flere mappinger.

5 Oppsummering og konklusjon

Prosjektet *Felles terminologi for klassifikasjon med Dewey* har utredet metoder for etablering av overganger fra Realfagstermer (og indirekte TEKORD) til Dewey. Deweymaterialet har begrenset seg til gruppene 500 og 600-640 i den uferdige, norske oversettelsen av DDC23.

Fellesvokabularet til Realfagstermer og TEKORD (TORT) med 3161 begrepspar som ble matchet i vårt forrige prosjekt, har vært et prioritert datagrunnlag for mappingarbeidet mot Dewey. En viktig del av arbeidet vårt har derfor vært å kvalitetssikre TORT gjennom opprettelse av en ny SKOS-fil med stabile identifikatorer, og gjennom intellektuell vurdering av relasjonstypene etter ISO 25964.

Realfagstermer har vært kildevokabularet vårt for mappingen mot Dewey, men gjennom å etablere mappinger fra begreper som også har mappinger til TEKORD, kan Realfagstermer potensielt brukes som mellomledd for overganger mellom TEKORD og Dewey, eller UDC til Dewey, siden TEKORD allerede er mappet til UDC.

Vi har kommet fram til en semiautomatisk metode bestående av automatisk genererte mappingforslag og intellektuell vurdering av disse basert på ISO 25964 og erfaringer fra CrissCross-prosjektet, som vi har brukt som utgangspunkt for vårt utkast til arbeidsbeskrivelse for mapping (se side 23). De to komponentene i metoden må spille på lag, og vi har derfor utviklet verktøyet μmapper som støtte for brukeren som skal gjøre den intellektuelle vurderingen. Vi har prioritert dette fordi vi har sett at det intellektuelle arbeidet er meget krevende, og at en trenger all støtte en kan få for å gjøre det effektivt.

Vi har generert 2804 automatiske mappingforslag fra Realfagstermer til DDC gjennom tekstsammenlikning og statistisk sammenlikning. Begge metodene har vist seg å være nyttige, men implementasjonene våre av begge har vært primitive, og vi ser mange muligheter for å utvikle dem videre.

- Tekstsammenlikningsmetoden har vært begrenset til eksakt strengmatching mellom begrep i Realfagstermer og klassebetegnelse i Dewey. Metoden ga 1681 forslag til mappinger mellom Realfagstermer og DDC (mappekandidater). 673 av Realfagstermene tilhører fellesvokabularet TORT, og gir derfor indirekte mappinger mellom UDC og DDC.

- Statistisk sammenlikning basert på korrelasjon mellom emneord og DDC i katalogdata fra BIBSYS med dokumenter klassifisert ved UBO fra januar 2012 til april 2014 der det finnes minst en Realfagsterm og en UBO-DDC-klassifikasjon. Metoden ga 1169 forslag til

mappinger. 350 av Realfagstermene tilhører fellesvokabularet TORT, og gir derfor indirekte mappinger mellom UDC og DDC.

Den maskinelle tilnærmingen har slik gitt oss et utgangspunkt for å starte mappingen, i form av en god del mappekandidater som behandles videre enten ved stikkprøver eller med mer omfattende kontroll.

For at overgangen fra Realfagstermer til DDC skal bli kunne bli et nyttig hjelpemiddel for sluttbrukere, har vi en lang vei å gå, både når det gjelder å gjøre den stor nok og når det gjelder å kvalitetssikre den. Både metodene og verktøyet μmapper har forbedringspotensial og vil være gjenstand for videre diskusjoner og utvikling i UBO-prosjektet *På vei mot en generell norsk tesaurus*[17] som skal utrede metoder for mapping av HUMORD mot DDC. En eventuell videreføring av dette prosjektet, samt utfallet av forprosjektet til Nasjonalbiblioteket og Universitetsbiblioteket i Oslo for å etablere en universell norsk tesaurus, vil være avgjørende for veien videre for en eventuell mapping av Realfagstermer/TEKORD mot Dewey og for innholdet i det videre samarbeid mellom Realfagstermer og TEKORD.

[17] http://www.ub.uio.no/om/prosjekter/tesaurusprosjektet/

Arbeidsbeskrivelse for mapping Realfagstermer/TEKORD/DDC

Denne arbeidsbeskrivelsen gjelder det intellektuelle arbeidet med mappingen som blant annet omfatter bruk av kilder, vurdering av betydningsomfang og relasjonstyper. Mappingen gjøres i µmapper-verktøyet. Beskrivelsen må betraktes som foreløpig.

Arbeidsbeskrivelsen er basert på ISO-standardens del 2 [3]og på CrissCross-prosjektets veiledning for mapping [5].

Prinsipper for mapping

Enveismapping. Realfagstermer er kildevokabular og norsk WebDewey (DDC) er målvokabular. Når det gjelder TORT mappet mot DDC, er begrepet i Realfagstermer utgangspunkt for mappingen.

En-til–en-mapping. Et enkeltbegrep kan mappes til flere ulike klasser i DDC. Det etableres altså flere enkeltstående relasjoner mellom begrepet og de passende DDC-klassene.

Relasjonstyper

EQ Eksakt samsvar: En samsvarsrelasjon mellom to begrep innebærer at de to begrepene kan bytte plass med hverandre, uten at det går ut over betydningen. Dette tilsvarer CrissCross' Determiniertheitgrad 4 (D4).

~EQ Tilnærmet eksakt samsvar. Typiske tilfeller: Begrepene kan samsvare i noen tilfeller, men ikke i andre. Begrepene kan ha overlappende betydning, eller små forskjeller i konnotasjon. Når man mapper mellom et klassifikasjonsskjema og en tesaurus er det vanlig å finne en klasse der klassebetegnelsen er lik en foretrukken term, men der nærmere undersøkelser viser at de to ikke er eksakt like.

BM Hierarkisk relasjon: Kildebegrepet er smalere enn målbegrepet.

NM Hierarkisk relasjon: Kildebegrepet er videre enn målbegrepet.

RM Assosiativ relasjon: Kildebegrepet kan assosieres med målbegrepet på en slik måte at man kan anta at målbegrepet vil være relevant for den som søker etter kildebegrepet..

Vurdring av mappekandidater

A. Begrepet som skal mappes har fått kun en foreslått kobling til DDC

1. Velg arbeidsliste i µmapper

2. Velg en foreslått relasjon

3. Bestem begrepets betydningsomfang

- Definisjon?
- Intern note for bruk?
- Emneordets relasjoner (synonymer/se også-henvisninger)
- Faktisk bruk på poster i biblioteksystemet
- Eventuelt sammenlikn med TEKORD

4. Bestem den forslåtte Deweyklassens betydningsomfang ved oppslag i norsk WebDewey og/eller amerikansk WebDewey

- Klassebetegnelsen
- Noter
- Registertermer
- Hierarkisk og faglig kontekst
- Faktisk bruk

5. Bestem relasjonstypen(e) og lagre.

B. Begrep som har flere enn én foreslåtte koblinger til DDC

1. Som for A (en relasjon), men vurdér og sammenlikn de foreslåtte klassenumrene med hensyn til faglig-/hierarkisk kontekst. Bestem ut fra emneordets betydning og faktiske bruk hvilken/hvilke klasse(r) det kan mappes til. Vi mapper ikke til en faglig kontekst/hierarki som vårt emneord ikke brukes i.

2. Bestem relasjonstypen(e) og lagre.

C. Foreslåtte retningslinjer for vanskelige tilfeller (diskuteres)

Her-noter

Her følger vi CrissCross og sier at emneord som står i her-noter ikke kan være D4, dvs. EQ. Disse kan ofte settes til ~EQ, men andre relasjonstyper bør også vurderes for hvert enkelt tilfelle.

Inkluderer-noter

Et emne som står i inkluderer-note mappes ikke som EQ eller ~EQ (CrissCross D4 eller D3). Vurdér om relasjonen er hierarkisk (BM eller NM).

Avslå relasjon

[Flere tilfeller vil komme til her etter hvert.]

D. Godkjenning av relasjoner

Minst to personer behandler hvert forslag. Etter at en relasjonstype har blitt bestemt, må den godkjennes av en annen person. Hvis en person A begynner med å bestemme relasjonstypen, vil relasjonen få status «venter på godkjenning» inntil en annen person B har godkjent relasjonen. Hvis person B velger ikke å godkjenne den, men i stedet endrer relasjonstypen til en ny verdi, vil status forbli «venter på godkjenning» inntil en annen person enn B godkjenner relasjonen.

Referanser

1. Kuldvere, L.V., et al. *Realfagstermer og TEKORD. RDF som plattform for sammenlikning og sammenføyning av emnesystemer? Rapport*. 2013, Oslo: [Forfatterne]. http://urn.nb.no/URN:NBN:no-36333

2. *Bibliografisk og emnemessig beskrivelse av UBOs samlinger : rapport fra en prosjektgruppe*, 2010, UiO: Universitetsbiblioteket: Oslo.

3. 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.

4. International Organization for Standardization *ISO 25964-1. Information and documentation : Thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval*. 2011, Geneve: ISO.

5. DFG-Projekt CrissCross *Leitfaden zur Vergabe von DDC-Notationen an SWD-Schlagwörtern*. 2010, Köln: Fachhochschule Köln. http://linux2.fbi.fh-koeln.de/crisscross/CrissCross_Endg_Grundlagenpapier_Sept2010.pdf

6. Hjortsæter, E. *Emneordskatalogisering: innholdsanalyse, emnerepresentasjon og lagring*. 2005, Oslo: Høgskolen i Oslo, Avdeling journalistikk, bibliotek- og informasjonsfag.

7. Vizine-Goetz, D., *Popular LCSH with Dewey numbers: subject headings for everyone.* Journal of Library Administration, 2001. **34**(3-4): s. 293-300.