På vei mot en generell norsk tesaurus?

Delprosjekt *Metodikk for mapping av Humord mot WebDewey*

Rapport

Are Gulbrandsen, Dan Michael O. Heggø, Unni Knutsen, Grete Seland

Oslo, 1. mars 2015

Innledning

Prosjekttildeling

Prosjektet *På vei mot en generell norsk tesaurus?* fikk støtte fra Nasjonalbiblioteket i 2014. Prosjektet er delt i to, separate aktiviteter:

1. Et forprosjekt som skal gi beslutningsgrunnlag i spørsmålet om det skal etableres en universell tesaurus med utgangspunkt i Humord
2. Å utvikle metodikk for mapping av tesaurusen Humord mot WebDewey

Forprosjektet (punkt 1) er et samarbeidsprosjekt mellom Nasjonalbiblioteket (NB) og Universitetsbiblioteket i Oslo (UBO). Det er utarbeidet en egen rapport for dette arbeidet.

Foreliggende rapport dekker dermed mappingmetodikkdelen av prosjektet (punkt 2).

I tildelingsbrevet fra Nasjonalbiblioteket[1] er aktivitetene i mappingprosjektet beskrevet slik: «Utvikle metodikk for mapping fra tesaurus til WebDewey, med Humord som utprøvingsarena. Utføres av Universitetsbiblioteket (UBO), gjerne sammen med andre fra Humord-samarbeidet.»

Mappingmetodikkprosjektet bygger på to tidligere NB-støttede prosjekter i regi av Realfagsbiblioteket ved UBO. I det første prosjektet (Kuldvere et al., 2013) ble det gjort sammenlikninger av terminologi mellom Realfagstermer og NTNUs tesaurus TEKORD. Ved hjelp av automatisk tekstsammenlikning fant man et fellesvokabular som utgjorde ca. 20 prosent av hvert av vokabularene. I en videreføring av prosjektet (Kuldvere et al., 2014) ble det utviklet metoder for mapping mellom Realfagstermer (inkludert fellesvokabularet med TEKORD) til aktuelle deler (500-gruppen og 600–640) av den foreløpige norske oversettelsen av Dewey Decimal Classification (DDC).

Det ble i denne forbindelse utviklet en prototype på et webbasert mappingverktøy, µmapper[2], for å gjøre prosessen med intellektuell vurdering av automatisk genererte mappingforslag smidigere enn en arbeidsprosess basert på regneark. Hovedmålet i inneværende prosjekt har vært å videreutvikle metodikk for datastøttet intellektuelt arbeid for mapping. 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.

Kort om vokabularene

Humord er en tesaurus som opprinnelig (oppstart 1993/1994) ble utviklet innenfor humaniora. Fra 2011 er vokabularet utvidet med termer fra samfunnsfagene. Humord omfatter i dag ca. 18 500 hovedtermer og ca. 8 500 se-henvisninger. Humord er helt fra starten av utviklet som et samarbeidsprosjekt. Indekseringssamarbeidet består i dag av universitetsbibliotekene i Oslo, Bergen og Tromsø samt Senter for studier av Holocaust og livssynsminoriteter. Humordarbeidet samordnes av en koordineringsgruppe og ledes av en Humord-koordinator fra UBO.

---

[1] https://www.ub.uio.no/for-ansatte/om-ubo/prosjekter/tesaurus-mapping/delte-dokumenter/tildelingsbrev-tesaurus.pdf

[2] https://github.com/scriptotek/mumapper/ (uttales: mymapper)

Realfagstermer er et kontrollert, pre-koordinert emneordsvokabular som hovedsakelig dekker fagområdene naturvitenskap, matematikk og informatikk. Vokabularet omfatter ca. 15 000 hovedtermer og ca. 2 000 se-henvisninger.

Begge vokabularer består av termer på norsk språk (bokmål), men termer på andre språk (eksempelvis engelsk, latin) forekommer også i Realfagstermer.

Hva er mapping og hvorfor er mapping nyttig?

Mapping kan defineres som en aktivitet hvor det etableres relasjoner (eller mappinger) mellom begreper i to ulike kontrollerte vokabularer. Settet av relasjoner mellom to vokabularer kalles gjerne en overgang (crosswalk) mellom de to vokabularene.

Overganger, enten mellom to emnevokabularer eller mellom et emnevokabular og et klassifikasjonsskjema, gir mulighet for søk på tvers av disse. Lokalt vil en overgang mellom Humord/Realfagstermer og Dewey åpne for muligheten til å søke med for eksempel en Humord-term og få treff også i dokumenter som ikke er indeksert med Humord, men klassifisert etter Dewey eller indeksert med Realfagstermer.

Internasjonalt er flere store emneordsvokabularerer som Library of Congress Subject Headings (LCSH), Gemeinsame Normdatei (GND) og Nuovo soggettario, allerede mappet til DDC. Ved å mappe UBOs vokabularer mot DDC, kobler vi oss dermed i praksis også mot et felles «nav» som dermed muliggjør flerspråklige søkeinnganger til både egne og andres samlinger. Dette bør kunne åpne for nye og spennende sluttbrukertjenester.

Merk at vi gjennom rapporten omtaler deweysystemet på ulike måter (Dewey Decimal Classification, DDC, Deweys desimalklassifikasjonssystem, Dewey). Vi omtaler også webverktøyet WebDewey synonymt med klassifikasjonssystemet.

Det teoretiske utgangspunktet for mappingarbeidet vårt er ISO-standard 25964 om tesauri (International Organization for Standardization, 2009, 2013). Det er særlig del to, *Interoperability with other vocabularies* (ISO 25964-2), som er relevant for vårt arbeid.

Prosjektdeltakerne og aktiviteter i prosjektet

Deltakere i mappingprosjektet har vært: Are Gulbrandsen, Dan Michael O. Heggø, Berit Sonja Hougaard, Viola Kuldvere, Vibeke Stockinger Lundetræ (fra 1. juni 2014), Mari Lundevall, Grete Seland (fra 15. januar 2015) og Unni Knutsen (prosjektleder).

Prosjektgruppens medlemmer har til sammen hatt 10 felles prosjektmøter og en rekke mindre arbeidsmøter. I tillegg til interne møter i prosjektet har det underveis vært kontakt med deweyredaksjonen i Nasjonalbiblioteket v/Elise Conradi og også med Språkbanken i NB. Prosjektmedarbeiderne Are Gulbrandsen, Dan Michael O. Heggø, Unni Knutsen og Mari Lundevall var til stede på EDUGs (European DDC Users Group) årlige møte 2014 i Reykjavik og presenterte prosjektet og mappingproblematikk på 1st Annual EDUG Mapping Working Group Meeting.

Mappingprosjektet er også presentert i Nasjonalbibliotekets skriftserie *Bibliotheca Nova*, 4-2014 (Knutsen & Gulbrandsen, 2014).

Økonomiske disposisjoner

I vedlegget til rapporten gis det en regnskapsoversikt per ultimo februar 2015. Den tildelingssummen vi mottok fra Nasjonalbiblioteket var støtte til begge delprosjektene nevnt

over. Vi har utelukkende brukt midlene på delprosjekt mapping. Utgiftene i prosjektet er knyttet til lønn til systemarkitekt (Are Gulbrandsen), til noe frikjøp av Dan Michael O. Heggø og fra 15. januar 2015 også til lønnsmidler til Grete Seland. Vi har også disponert noen midler i forbindelse med EDUGs konferanse i Reykjavik. Siden ekstern tilsetting av systemarkitekt fant sted noe inn i prosjektperioden, har vi fremdeles noen midler til disposisjon. Vi planlegger disse midlene brukt til lønn og aktiviteter fram mot neste bevilgning fra Nasjonalbiblioteket.

Rapportens struktur

Rapporten er strukturert som følger:

Først presenteres ISO-standard 25964-2 om interoperabilitet mellom vokabularer. Denne standarden danner det teoretiske og metodiske fundamentet som arbeidet vårt hviler på. Vi redegjør her for overordnet valg av mappingmodell, presenterer de ulike mappingrelasjonene og gjør rede for hvilke overordnede valg vi har tatt basert på anbefalingene i standarden og vokabularenes natur. Forholdet mellom standarden og SKOS[3] berøres.

I kapittelet om intellektuelle utfordringer i forbindelse med mapping presenteres både generelle betraktninger om mapping og mer spesifikke vurderinger av hva det innebærer å mappe tesaurusen Humord mot Dewey. Valg av mappingkandidater og mappingrelasjoner problematiseres. For å avdekke kompleksiteten som vil oppstå i forbindelse med den faktiske mappingen, foretar vi i øyeblikket en rekke testmappinger. Det knyttes også refleksjoner til nytten av et mappingverktøy når mappingkandidater skal velges og mappingrelasjoner uttrykkes.

Som et ledd i målet om dataassistert hjelp ved mapping, presenteres mappingsverktøyet ccmapper. ccmapper bygger videre på den tidligere nevnte µmapper som er utviklet av Realfagsbiblioteket. I dette kapittelet går vi gjennom hvilke datakilder som vi legger til grunn når vi ønsker å legge til rette for gode mappingkandidater. SKOS får en nærmere presentasjon. Ulike metoder som termvekting, lemmatisering m.v. presenteres.

Avslutningsvis trekkes det linjer framover mot det arbeidet som skal utføres i neste prosjekt.

[3] Simple Knowledge Organization System http://www.w3.org/2004/02/skos/

ISO-25964-2 og SKOS som utgangspunkt for mapping

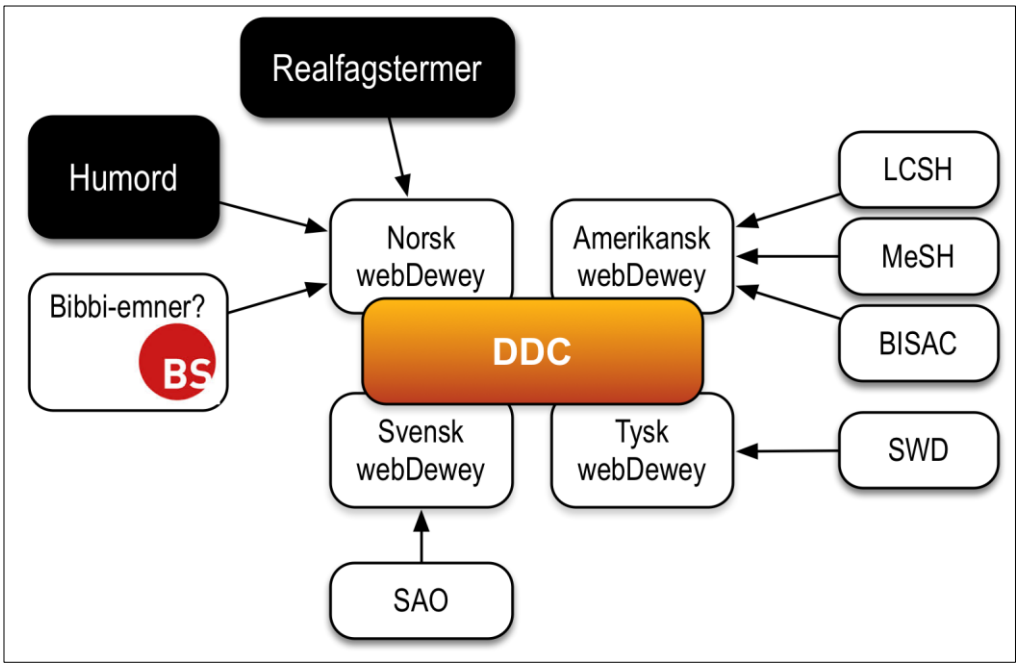
Som nevnt innledningsvis danner ISO-standard 25964, (International Organization for Standardization, 2009, 2013) det teoretiske og metodiske fundamentet for vårt mappingarbeid. ISO standarden er i to deler. ISO 25964-1 beskriver hovedsakelig hvordan en tesaurus er bygd opp og gir anvisninger om beste praksis på området. ISO 25964-2 *Interoperability with other vocabularies* springer ut fra en erkjennelse av at det er et stort behov for å identifisere og lokalisere relevant informasjon på tvers av store samlinger. Dette danner et behov for å skape semantisk interoperabilitet. I gjenfinningsøyemed er målet med interoperabilitet mellom vokabularer å koble en term fra ett vokabular til en tilsvarende term i ett eller flere andre vokabularer. I ISO 25964-2 gis det grunnleggende beskrivelser av andre typer vokabularer (som klassifikasjonssystem) og anbefalinger om mapping mellom disse og tesauri. Av de to delene av standarden er det derfor ISO 25964-2 som har hatt størst betydning for vårt arbeid. Når vi videre i rapporten viser til ISO-standarden menes implisitt ISO 25964-2.

ISO 25964-2 definerer verbet map slik: «establish relationships between the concepts of one vocabulary and those of another», mens mapping defineres som «process of establishing relationships between the concepts of one vocabulary and those of another» (International Organization for Standardization, 2013, s. 7).

Valg av strukturell modell

Standarden beskriver innledningsvis ulike strukturelle modeller for hvordan vokabularer kan mappes sammen. Standarden legger vekt på at man i starten av alle mappingprosjekter må klargjøre hvilken modell som skal brukes og hvilken retning mappingen skal ha. Det gis gode råd for valg av strukturell modell.

I vårt tilfelle var det naturlige valget å satse på nav-modellen (figur 1) hvor målet er å mappe våre vokabularer (Humord og Realfagstermer) mot Deweys klassifikasjonsssystem (den norske oversettelsen). Når Humord mappes mot DDC og Library of Congress Subject Headings (LCSH) er mappet mot samme kilde, ser vi ikke behovet for å bruke store ressurser på å også mappe Humord direkte mot LCSH. I et framtidig sluttbrukersystem kan vi se for oss at brukeren ved å søke på en term i Humord kan videresøke inn i baser og tjenester med LCSH-emneord. Den underliggende koblingen vil være et deweynummer. Svakheten ved denne metoden er at et deweynummer ikke alltid kun rommer ett emne, men flere. En kobling basert på mapping via et klassifikasjonsnummer vil dermed i visse tilfeller medføre gjenfinning av noen irrelevante dokumenter. Vår vurdering er at siden DDC har en høy grad av granularitet og i vesentlig grad sammenstiller emner som er beslektede, vil dette ikke skape et vesentlig problem for brukeren.



Figur 1: Dewey Decimal Classification (DDC) som nav. Humord og Realfagstermer er her mappet til Norsk webDewey. Andre vokabularer er mappet til DDC på tilsvarende måter.

Som vist i figur 1 er det å bruke DDC som nav, en strategi som er utbredt internasjonalt. ISO-rapporten anbefaler denne framgangsmåten på «the reconciliation of vocabularies that have been independently developed and/or have already been applied to collections». (s. 20)

Når det gjelder retningen på mappingen opererer ISO-standarden med begrepene source vocabulary (kildevokabular) og target vocabulary (målvokabular). Kildevokabularet defineres som «vocabulary that serves as a starting point when seeking a corresponding term or concept in another vocabulary» (s. 12), mens målvokabularet er definert som «vocabulary in which a term or concept is sought corresponding to an existing term or concept in a source vocabulary» (s. 13). I vårt prosjekt er Humord kildevokabularet som mappes inn mot målvokabularet DDC.

Det er viktig å være klar over at vi mapper betydningsinnholdet i ordene. Det betyr at vi må være bevisst på at sammenfallende ord ikke nødvendigvis bærer samme betydning.

Ulike typer mappingrelasjoner

Når to termer fra ulike vokabularer skal mappes, må man etablere hva slags type relasjon det er mellom dem. Mappingrelasjonene baserer seg på relasjonstyper som er kjent fra tesauruskonstruksjon, med ekvivalens-, hierarkiske- og assosiative forbindelser. ISO-standarden opererer med to typer ekvivalens. Den hierarkiske relasjonstypen går begge veier («broader» og «narrower»). Vi får derfor følgende fem relasjoner:

=EQ	(EQ = Equivalence). 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

En type ekvivalensrelasjon er «compound equivalence mapping» hvor et begrep i ett vokabular tilsvarer flere begreper i et annet vokabular. Man kan da bruke en kombinasjon av mappingrelasjonene og boolske operatorer (AND og OR, symbolisert ved + og │). Et av eksemplene i ISO-standarden (s. 36) viser hvordan tesaurustermen «institutions» kan mappes til tre aktuelle klassenumre: «institutions EQ E100 │ H100 │ D100». Alternativt kan man bruke flere uavhengige mappinger. Vi planlegger å bruke sistnevnte løsning, og unngår dermed bruk av boolske operatorer.

«Compound equivalence mapping» kan også være aktuelt i de tilfellene hvor det ene vokabularet bruker sammensatte termer der det andre vokabularet splitter. Begrepet «kvinnelige ledere» i vokabular 1 kan eksempelvis tilsvare «kvinner» + «ledere» i vokabular 2. Testmappingen vi har utført tilsier at det ikke er behov for å ta i denne løsningen, fordi Humord i stor grad unngår sammensatte termer.

ISO 25964-2 og SKOS

Et av hovedmålene i alle våre utviklingsprosjekter har vært å publisere både vokabularene og mappingene våre med tanke på den semantiske weben. Av den grunn har vi helt fra starten av ønsket å bruke SKOS/RDF-standardene i arbeidet vårt.

SKOS-modellen opererer med *closeMatch*, *exactMatch*, *narrowMatch*, *broadMatch* og *relatedMatch*. I tillegg har modellen noter som er nyttige for definisjoner, redegjørelse for hvordan begrepet er brukt, historikk, kommentarer med videre. Til tross for at SKOS er en betydelig forenkling av ISO 25964-1-modellen, akkomoderer den på mange måter en tesaurus godt og den inkluderer alle relasjonene vi har behov for å uttrykke. Det er selvsagt ikke tilfeldig at ISO-standarden og SKOS sammenfaller langt på vei. De to utviklingsmiljøene har fulgt hverandre tett. I ISO-standardens del to (s. 43) er dette uttrykt slik:

> As yet there is no standard schema that fully complies with this part of ISO 25964, and the development of such a schema is not within scope. However, this is a rapidly evolving field and implementers of this part of ISO 25964 should be alert to developments among interested parties, e.g. the SKOS user community.
>
> The schemas developed for storage purposes may also be used or adapted to enable publication of the mappings. A SKOS-compliant format […] is recommended if use in the Semantic Web is desired.

Dette understøtter valgene våre av SKOS og RDF.

Overordnede beslutninger

ISO-standarden (s. 31) gir anvisninger for hvilke beslutninger som må tas ved starten av ethvert mappingprosjekt. Under listes spørsmålene som stilles og vår respons til disse:

Which overall model or combination of models to use:

Som redegjort for under "Valg av strukturell modell", har vi landet på en modell[4] der Humord er kildevokabular og Dewey er målvokabular. DDC er navet i denne modellen. I første omgang modelleres Humord ved hjelp av SKOS og begrepene mappes til deweyklasser.

How much to differentiate the mappings, in the following respects:

- whether to distinguish between equivalence and other types of mapping such as hierarchical and associative

Vi ønsker å bruke alle de tre mappingrelasjonstypene nevnt over.

- whether to accept compound equivalence mappings:

Som argumentert for over kommer vi ikke til å benytte compound equivalence mappings.

- whether to distinguish between exact and inexact equivalence;

Vi ønsker å skille mellom =EQ og ~EQ.

- whether to enable establishment of more than one mapping per concept;

Siden vi mapper fra en tesaurus mot et faglig inndelt skjema anser vi det som relevant å mappe ett og samme Humord inn mot flere plasseringer i Dewey i form av flere uavhengige mappinger.

Whether and how to enable human mediation in the conversion process

I forbindelse med forrige NB-støttete prosjekt ved Realfagsbiblioteket ble det utviklet en prototype på et mappingverktøy, µmapper for å gjøre mapping mellom Realfagstermer og DDC enklere og mer effektivt. Vi ønsker å bygge videre på erfaringene som er gjort i dette prosjektet slikt at det blir enklere for den som skal foreta mappingene å fatte en riktig beslutning om mappingkandidater og relasjonstyper.

[4] I ISO-standarden (s. 19) omtales dette som «Model 3».

Intellektuelle utfordringer i forbindelse med mapping

Mapping som kunnskapsorganisatorisk fenomen

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? Det er slike problemstillinger vi bryner oss på i forbindelse med mappingarbeid. Å sette seg som mål å etablere koblinger mellom begreper fra to ulike vokabularer, er et ambisiøst foretagende. Som i all kunnskapsorganisatorisk praksis, er man ved mapping nødt til å etablere et skarpere skille mellom begrepskategorier enn det man finner i naturlig språk.

Et typisk trekk ved naturlig språk, er at det er gjennomsyret av flertydighet (polysemi) – der samme term har flere beslektede betydninger (som for eksempel når «horn» både betegner utvekster på dyrehoder, et bakverk og et musikkinstrument). Vi finner også mange tilfeller av nærsynonymi (kvasisynonymi), der ulike termer har tilgrensende betydninger (som for eksempel «gjennomskinnelighet» og «gjennomsiktighet»). Glidende betydningsskiller i naturlig språk behandles i tesauri ved etablering av skiller mellom foretrukne og ikke foretrukne termer (f.eks. hvis «Norden» etableres som foretrukken term med «Skandinavia» som henvisning). I et klassifikasjonsskjema er det mest framtredende eksemplet på etablering av skarpe skiller mellom begrepskategorier, det at man deler kunnskapsuniverset inn etter en titalls-struktur.

Eksemplene over viser at man i kunnskapsorganisasjon har erfaring med å finne praktiske løsninger for å håndtere kompleksiteten i naturlig språk. I forbindelse med mapping dukker det opp ytterligere utfordringer. Ikke bare skal man koble vokabularer som har hvert sitt inventar av termer – man må også forholde seg til at det er brukt ulike inndelingskriterier ved etableringen av de hierarkiske strukturene. Betydningsskillene kan altså være håndtert ulikt i hvert av vokabularene. For eksempel kan «gjennomskinnelighet» og «gjennomsiktighet» være etablert som to foretrukne termer i det ene vokabularet (med se også-henvisning mellom seg) – og altså være betraktet som to ulike begreper – mens det andre vokabularet kan håndtere samme fenomen som nærsynonymi (der ett begrep er representert av to termer, hvorav den ene er foretrukken og den andre er en henvisning).

Selv om formålet med mapping er å koble begreper, er vi avhengige av deres representasjoner i form av termer. Den hierarkiske konteksten som en term inngår i, bidrar til å klargjøre hvilken betydning man i et gitt vokabular tilskriver en term. Siden det kan være brukt ulike kriterier ved opprettelsen av disse strukturene, kan man ikke uten videre koble termer i ulike vokabularer på grunnlag av termlikhet. Dessuten vil termer som er forbundet med hverandre i ett vokabular, kunne være spredt på ulike steder i strukturen i et annet vokabular. Ambisjonen om mapping mellom ulike kunnskapsstrukturer kan synes umulig i sin natur, med mindre man inntar en pragmatisk holdning. Denne er motivert av ønsket om å tilby sluttbrukere bedre emnemessig tilgang til ressursene i katalogen.

Ved mapping av en flat emneordsliste mot et annet vokabular (for eksempel slik det ble gjort i forbindelse med Realfagstermer mot Dewey), slipper man frustrasjonene man møter i forbindelse med to kolliderende hierarkiske strukturer. På den annen side har man da ikke den hierarkiske konteksten å lene seg på i tolkningen av termene i kildevokabularet. Ved etablering av en mappingmetodikk i dette prosjektet, er utgangspunktet Humord. Planen er likevel at metoden skal være anvendelig også for kobling av andre vokabularer mot WebDewey.

Mapping mellom ulike strukturer: Tesaurus kontra klassifikasjonsskjema

En iøynefallende utfordring knyttet til mapping mellom en tesaurus og et klassifikasjonsskjema, er at man står overfor to forskjellige overordnede inndelingsprinsipper. Mens en tesaurus inndeles etter *emne,* deles et klassifikasjonsskjema inn etter *fag*. Dermed vil et emne som er samlet ett sted i en tesaurus, bli spredt på håndteringen av dette emnet i forbindelse med ulike fagområder i et klassifikasjonsskjema. I en tesaurus forholder man seg til et inventar av termer i en hierarkisk struktur. I et klassifikasjonsskjema som Dewey finner man i tillegg også et stort noteapparat, hjelpetabeller og retningslinjer for bruk. Dette er utfordrende både i forbindelse med utvalgskriterier for mappingkandidater i dataverktøyet, og for den som skal yte det intellektuelle bidraget i selve mappingen.

På bakgrunn av disse observasjonene, er det naturlig å forvente at det i vårt prosjekt vil bli mange mappinger der ett begrep i Humord får to eller flere uavhengige mappinger mot WebDewey, fordi begrepet vil være spredt på ulike faglige plasseringer. Det vil også være aktuelt med mapping mot nummerspenn, for eksempel når den generelle og den spesifikke håndteringen av et emne er fordelt på sidestilte klassenumre. Dette ser vi f.eks. når man skal koble emneordet «samlinger (bibliotek)» i Humord mot WebDewey, der dette emnet vil være fordelt på klassenumrene 026 og 027, uten et felles overordnet nummer.

I det tilfellet at man har sammensatte Humord-emner, vil vi mappe mot bygde numre der det er mulig, for eksempel når termen «Dakota folkediktning» kobles til 398.2089975243. Såkalt «compound equivalence mapping» – dvs. tilfeller der ett Humord-emne tilsvarer sammensetningen av flere klassenumre i hovedtabellen som ses avhengig av hverandre – planlegger vi som nevnt tidligere ikke å etablere. Sammensatte emner som regel lar seg klassifisere enten på et hovednummer eller ved bygde numre. Dessuten fraråder ISO-standarden denne formen for mapping fordi det resulterer i støy.

En mulig innfallsvinkel i forbindelse med mapping av emner som dekkes av flere klassenumre i Dewey, kan være å mappe mot numrene for oversiktsverker og tverrfaglig plassering. Mange emner har i Dewey angitt et klassenummer for generell behandling, i tillegg til spesifikke anvendelser plassert på tilgrensende numre. Et eksempel på dette er når «dyr» klassifiseres på 571.1 (biologi) med undernumre, mens «oversiktsverker om dyr» skal settes på 590. Man kan da vurdere å mappe kun mot klassenummeret for oversiktsverker.

For emner som opptrer i flere hovedklasser i forbindelse med ulike faglige plasseringer, vil det ofte være angitt i tabellen hvor man skal plassere tverrfaglige verker. Et emne som døden vil forekomme i forbindelse med blant annet filosofi, religion, medisin og folkeminne, men har også angitt et klassenummer for tverrfaglige verker, innen sosiologi på 306.9. Mapping mot klassenummeret for oversiktsverker og tverrfaglig håndtering peker seg ut som en god pragmatisk løsning, men er likevel ikke uproblematisk. Emnets kontekst i Humord kan tilsi en annen plassering – for eksempel befinner Humord-emnet «døden» seg i hierarkiet «Helse > Kroppen > Døden», ikke under sosiologi. Samtidig vil man se at anvendelsen av emnet i indeksering ikke begrenser seg til dette aspektet.

En annen utfordring i mapping fra tesaurus til klassifikasjon, angår distinksjonen mellom *postkoordinerte* og *prekoordinerte* systemer: Humord-tesaurusen er beregnet for indeksering med emneord for postkoordinert gjenfinning. Deweyklassifikasjonen på sin side, er beregnet for prekoordinert indeksering. Dette byr på store utfordringer. De frittstående emneordene i Humord vil få hver sin mapping mot aktuelle deweyklasser. Ved indeksering med bruk av Humord, får hvert dokument gjennomsnittlig tildelt fire frittstående emneord som vil være mappet mot ulike klassenumre. Et dokument om «sosiobiologisk etikk» vil blant annet være

tildelt Humord-emnene «sosiobiologi» og «etikk», som vil peke til hver sin hovedklasse. I et prekoordinert system med emneord i streng, ville man omvendt hatt «sosiobiologisk etikk», som kunne blitt mappet mot ett klassenummer. Det ser ut til at emneord i streng ville vært enklere å mappe mot Dewey enn frittstående emneord i en tesaurus, slik som i Humord.

Emneordet «morfologi» i Humord kan illustrere hvordan vi må bruke emneordenes kontekstuelle plassering til å forstå hvilken av flere mulige betydninger av et begrep vi har med å gjøre. I Humord finner vi «morfologi» som et underemne av «grammatikk». Dette hjelper oss til å avklare at det aktuelle klassenummeret i WebDewey vil være 415.9 under lingvistikk, og at det innenfor Humords kontekst ikke vil være aktuelt å lage koblinger til morfologi i forbindelse med menneskers eller dyrs anatomi, som vi finner på to helt andre steder i deweytabellen. Dermed kan vi etablere en ekvivalensrelasjon (eksakt samsvar) mellom emneordet «morfologi» og klassenummeret 415.9 – eller kan vi det? Hva med det forholdet at man ved hjelp av tilleggstallet -59 fra hjelpetabell 4 kan få uttrykt morfologi i forbindelse med det enkelte språks grammatikk, og dermed kan finne morfologi på utallige klassenumre? Hvis vi mapper mot bygde numre i andre sammenhenger (jfr. Dakota-eksemplet over), hvorfor skulle man da ikke gjøre det her? Skal det kun gjøres når emneordet i Humord er så spesifikt at man er nødt til å etablere et bygd nummer for å få emnet uttrykt i WebDewey, men ikke i de tilfellene hvor man har en generell plassering på et klassenummer (som for morfologi)? Kriterier for valg i slike situasjoner er fortsatt under vurdering.

Utelukkelsen av mapping mot anatomi-betydningene av «morfologi» er heller ikke uproblematisk med tanke på det overordnede målet om en generell norsk tesaurus – i den bør vel alle betydninger av begrepet som termen representerer forekomme? Hvis man på sikt mapper mange ulike vokabularer mot WebDewey, vil man måtte opprettholde den opprinnelige Humord-konteksten for å kunne forsvare en ekvivalensrelasjon fra «morfologi» til 415.9, ettersom denne relasjonstypen forutsettes å kunne gå begge veier mellom vokabularene.

Strukturelle likhetstrekk mellom Humord og faglig basert klassifikasjon

I tillegg til de ulike prinsippene som ligger til grunn for tesauri og klassifikasjonssystemer, må man forvente at ethvert kunnskapsorganiseringssystem vil bære preg av sin tilblivelseshistorikk. I vårt tilfelle ser vi for eksempel at Humord inneholder avvik fra tesaurusstandarden, og at Dewey er preget av valg som er foretatt i løpet av dets lange historie. Eksempler i forbindelse med Humord er at man har et avgrenset sett av topptermer som minner mye om en faglig inndeling – noe som er forståelig med tanke på Humords utspring i emneordsarbeid ved flere instituttbibliotek. Videre kan det forekomme forskjellige inndelingskriterier innen ett hierarki. I Dewey finner vi elementer av fasettert struktur (f.eks. biologi og musikk) i et hovedsakelig enumerativt skjema. Dette gjør arbeidet med å etablere kriterier for utvalg av mappingkandidater mellom Humord og WebDewey ekstra utfordrende. Vi kan ikke forutsette at et mønster som vi finner i forbindelse med testmapping av ett emneområde i Humord, vil fungere på samme måte et annet sted i vokabularet.

Spørsmålet om i hvilken grad man ved mapping skal ta hensyn til faktisk bruk av en tesaurusterm slik den framtrer i katalogdataene, kan også illustreres med emneordet «hester», som i Humord er plassert i følgende hierarki: Realfag > Naturvitenskap > Biologi > Zoologi > Dyreliv > Dyr > Vertebrater > Pattedyr > Hester. Likevel er dette emneordet brukt på katalogposter for dokumenter som handler om emner som motiv i kunsten. Man kan se for seg at i en allmenn tesaurus burde man ha flest mulig topptermer for å unngå slike dilemmaer – og emnene burde ikke vært innplassert i faglige hierarkier. I Humord finner vi den omvendte situasjonen, ved at alle de 18 500 hovedtermene er plassert innenfor 26 topptermer. Dette må

nødvendigvis skape store utfordringer. Strukturelt er Humord svært lik et klassifikasjonsskjema, men den anvendes som en tesaurus – med frittstående emneord til bruk i postkoordinert indeksering.

Dilemmaene som oppstår i forbindelse med Humords faglige orientering (emner plassert inn i faglige hierarkier) – i motstrid med prinsippet om emneinndeling i tesauri – kan belyses ytterligere ved hjelp av «tobakk». Dette emnet er et yndet eksempel når man skal illustrere hvordan et emne kan være spredt på ulike fagområder i et klassifikasjonsskjema. Emnet «tobakk» er i Dewey spredt over hele tabellen – på klassenumrene for botanikk, etikk, religion, landbruk, menneskelig toksikologi, produksjonsteknologi, skikker og smugling. I Humord finner vi emneordet «tobakk» plassert i følgende hierarki: Næringsliv og økonomi > Produksjon > Produkter > Nytelsesmidler > Tobakk. Tilsier dette at vi bør mappe begrepet kun mot plasseringen «tobakk» har i WebDewey i forbindelse med produksjonsteknologi?

Ofte har vi å gjøre med en uoverensstemmelse mellom struktureringsprinspipp og indekseringspraksis. «Tobakk»-eksemplet viser at man ved mapping må vurdere om man for kildevokabularet utelukkende skal ta stilling til dets struktur, eller også ta hensyn til faktisk bruk. Skal vi primært ta hensyn til at Humord har en overordnet faglig struktur (selv om det ikke burde vært slik ifølge prinsippene for tesauruskonstruksjon) – eller at den i faktisk indeksering brukes som en tesaurus (uten konsekvent hensyntagen til emneordenes faglige plassering i hierarkiene)? Hvilken konsekvens får avgjørelser forhold til disse spørsmålene med tanke på sluttbrukerperspektivet?

Hvis vi ser på katalogdataene for hvilke dokumenter som har blitt indeksert med emneordet «tobakk» i Humord, ser vi at de har svært ulike klassenumrene. Dette gjenspeiler den spredte faglige plasseringen dette emnet har i Dewey. Skal vi da mappe mot alle disse plasseringene? I morfologi-eksemplet over, brukte vi Humords domene som kriterium for å velge én mapping framfor potensielt tre. I tilfellet med tobakk, kan vi ikke si at Humords domene avgrenser seg til produksjonsaspektet av tobakk (selv om det er i dette hierarkiet termen er plassert). Skal vi da mappe mot alle plasseringene unntatt klassenumre innen realfag og medisin (som dekkes av Realfagstermer og MeSH) – altså et utvalg av aspektene som er listet over? I drøftingen av denne typen problemstillinger ser vi at Humords struktur med et lite antall topptermer, «tvinger» termer inn i hierarkier som ikke gjenspeiles i den bruken de ifølge katalogpostene faktisk har. Vi ser også at man ikke kan komme langt nok med statistisk mapping, og er nødt til å se på konteksten i både kilde- og målvokabularet.

Mappingkandidater: Valg av begrepspar og avklaring av relasjonstyper

Som tidligere nevnt opererer ISO 25964-2 med ekvivalensrelasjoner (eksakt og tilnærmet eksakt samsvar), hierarkiske (overordnet og underordnet) og assosiative relasjoner. I fortsettelsen brukes SKOS-betegnelsene for disse relasjonstypene: skos:exactMatch, skos:closeMatch, skos:broadMatch, skos:narrowMatch og skos:relatedMatch. La oss se på noen eksempler på hvordan mapping mellom Humord og WebDewey kan arte seg i praksis – først i forbindelse med vurdering av hvilke elementer det skal etableres mappinger imellom.

Vi så tidligere at emneordet «samlinger (bibliotek)» i Humord kan være aktuell å koble til nummerspennet 026-027, dvs. klassenumrene for samlinger i henholdsvis generelle bibliotek (027) og bibliotek innenfor bestemte fag og emner (026). Dette er eksempel på at Humord har en del allmenne termer som i Dewey ikke utgjør verken første eller andre inndelingskriterium, og derfor blir begrepet som representeres av emneordet i Humord spredt på flere steder i deweytabellen. Vi får da et tilfelle av mapping av ett begrep i Humord som har to uavhengige mappinger i WebDewey. Man kan enten lage en narrowMatch til nummerspennet 026-027,

eller lage en uavhengig narrowMatch til hvert av disse klassenumrene. Emneordet i Humord er mer generelt enn hver av disse, og dessuten har hvert av numrene det mappes mot videre underinndeling i WebDewey. Hvis man ikke skal foreta flere uavhengige mappinger av emneordet «samlinger (bibliotek)», vil en alternativ løsning være å etablere en broadMatch til 020 Bibliotek- og informasjonsvitenskap.

Hvis vi tar for oss emneordet «normativ etikk» i Humord, er det enkelt å finne en kobling til 170.44 i WebDewey som har klassebetegnelsen Normativ etikk. Den hierarkiske konteksten for termen og klassenummeret stemmer godt overens, så vi kan etablere en ekvivalensrelasjon – men bør den være en exactMatch eller en closeMatch? Hvis man i det hele tatt skal kunne bruke exactMatch ved mapping mot klassenumre, vil dette være et godt eksempel på et slikt tilfelle. Man må bare være klar over at aspekter av begrepsinnholdet tilordnet et klassenummer også kan være innbakt i andre klassenumre, for eksempel: Klassenummeret 170.44 gjelder for «oversiktsverker om normativ etikk», mens normativ etikk i spesifikke sammenhenger, vil komme på klassenummeret for det bestemte aspektet, for eksempel «yrkesetikk» på 174. Dette er en gjennomgående tematikk i forbindelse med mapping mellom en tesaurus og et klassifikasjonsskjema.

I samme hierarki som «normativ etikk» i Humord finner vi «sinnelagsetikk», som vi ikke finner i WebDewey. Her må man antakeligvis etablere en mapping mot 170.4, Spesielle emner innen etikk. Isolert sett vil dette borge for en broadMatch, fordi begrepet i målvokabularet er bredere enn i kildevokabularet. Samtidig må man velge å se bort fra at «sinnelagsetikk» (på grunn av manglende underinndeling i WebDewey) dermed får en Deweyplassering overordnet «normativ etikk», selv om begrepene i seg selv snarere er sideordnet og burde hatt klassenumre på samme nivå. Uansett viser dette eksemplet at når emneordet i Humord mappes mot WebDewey og blir tilgjengelige fra samme grensesnitt, vil de delene av Humord som har høyere granularitet enn deweytabellen, berike WebDewey med spesifikke termer som vil være nyttige både for sluttbrukere og i forbindelse med indeksering.

Emneordet «heraldikk» i Humord kan stå som eksempel på en stadig tilbakevendende problemstilling som vi har støtt på i prosjektet, nemlig valget mellom closeMatch kontra broadMatch. Termen «heraldikk» har følgende innførsel i Humord:

> Heraldikk
> BF Slektsvåpen
> BF Våpenskjold
> OT Kulturkunnskap
> TT Kulturkunnskap
> UT Emblemer
> UT Faner
> UT Flagg
> UT Riksvåpen
> SO Ordener (Utmerkelser)

I WebDewey finner vi «heraldikk» på klasse 929.6:

> 900 Historie og geografi
> 920 Biografier og genealogi
> 929 Genealogi, navn, insignier
> 929.6 Heraldikk
> Inkluderer: Våpenmerker; byvåpen, kommunevåpen
> Her: Våpenskjold

I mappingarbeidet må vi spørre oss om termen «heraldikk» i Humord representerer det samme begrepet som klasse 929.6 i WebDewey? Dette ser i utgangspunktet ut som en enkel exactMatch, siden Humord-emnet forekommer som egen klassebetegnelse i WebDewey. Det viser seg imidlertid å være flere kompliserende faktorer:

Hvis vi ser på den hierarkisk overordnede konteksten, er «heraldikk» i Humord emnemessig plassert under kulturkunnskap, mens det i klasse 929.6 i WebDewey er faglig plassert under «genealogi», som igjen ligger under historiefaget. I Humord finnes «genealogi» i et annet hierarki enn «heraldikk», nemlig som alternativ term til «slektsforskning» som vi finner under «historieforskning». Hvis vi ser på den hierarkisk underordnede konteksten, finner vi i Humord bl.a. «faner» og «flagg». I WebDewey finnes dette på 929.92, altså som et undernummer av en klasse som er sideordnet 929.6 hvor vi fant «heraldikk». Kan vi betrakte Humord-emnet «heraldikk» og klassenummer 929.6 som ulike representanter for samme begrep? I hvilken grad er de like, og hvilken relasjonsbetegnelse er det aktuelt å bruke her?

Som vi så i morfologi-eksemplet ovenfor, må vi – for at mapping i det hele tatt skal være mulig – bruke termenes kontekst for å avklare begrepsinnhold. Utfra Humord-konteksten forstår vi at det her er snakk om våpenmerker og symboler for slektskap. I WebDewey finner man også «heraldikk» i forbindelse med klassenummeret for «kongehus, adel og ridderordener» på 929.7, som har registerinnførsel på «Arverekkefølge (heraldikk)», så den anvendelsen kan vi se bort fra. Vi må også velge å se bort fra at de underordnede termene i Humord kan være plassert flere steder i WebDewey – for eksempel vil flagg og faner ikke bare finnes på 929.92, men også under militærvitenskap på 355.15.

At vi bruker Humord-konteksten for å forstå begrepsinnholdet til en term, innebærer derimot ikke at vi kan regne med at hierarkiet over og under en gitt term/klassenummer skal stemme overens – det ligger i sakens natur, når man mapper mellom en tesaurus og et klassifikasjonsskjema. Så her må vi velge å lage en kobling mellom emneordet «heraldikk» og klassenummeret 929.6 – men hvilken relasjon skal vi betegne den med? Konteksten i kilde- og målvokabularene utelukker en exactMatch, så vi står overfor valget mellom en closeMatch og en broadMatch. To av de underordnede termene til emneordet i Humord finnes ikke som oppslag i WebDewey («emblemer» og «riksvåpen»), mens de to øvrige ligger på et annet klassenummer («flagg» og «faner» på 929.92). Ingen av termene i inkluder-noten på 929.6 finnes som emneord i Humord. Likevel er dette åpenbart det aktuelle klassenummeret å bruke for denne mappingen. Siden vi ikke kan konstatere en generisk relasjon mellom begrepene i de to vokabularene, velger vi å etablere ekvivalensrelasjonen closeMatch.

Testmapping

For å avdekke kompleksiteten som vil dukke opp i forbindelse med den faktiske mappingen, foretar vi i prosjektet en rekke testmappinger. Eksemplene i de foregående avsnittene er hentet fra dette arbeidet. Erfaringene vi høster i forbindelse med testmappingen skal bygges inn som algoritmer i mappingverktøyet. Problemstillinger som ikke lar seg løse ved automatiserte metoder, må etableres som retningslinjer for konsekvent praksis i forbindelse med det intellektuelle bidraget i mappingarbeidet. I testmappingen søker vi først fram aktuelle mappingkandidater til et Humord-emne, og får fram hvordan ulike vurderingskriterier for valg av kandidater vil virke inn på hvilke mappinger man får (f.eks. valg av flere uavhengige mappinger av ett emne, kontra mapping mot klassenummeret for et overordnet begrep). Vi tar også stilling til hvilken relasjon hver mapping bør betegnes med. Her dukker det opp mange dilemmaer.

La oss illustrere med et eksempel: Humord-emnet «samarbeid» inngår i følgende hierarki: Samfunnsvitenskap > Sosiologi > Sosiale prosesser > Samarbeid, med de underordnede termene Gruppearbeid, Internasjonalt samarbeid og Regionalt samarbeid. Når vi mapper mot 302.14 som har klassebetegnelsen «sosial deltakelse» – skal vi da bruke en broadMatch eller en narrowMatch? Hvis vi ser på overordnet kontekst til Humord-emnet («sosiale prosesser», «sosiologi», «samfunnsvitenskap»), kan «samarbeid» tolkes som smalere enn «sosial deltakelse», og vi vil få en broadMatch fra Humord-emnet til klassenummeret. Hvis vi derimot betrakter den underordnede konteksten til Humord-emnet (dvs. Gruppearbeid, Internasjonalt samarbeid og Regionalt samarbeid), kan «samarbeid» tolkes som bredere enn «sosial deltakelse», og vi vil få en narrowMatch.

Første del av testmappingen dreier seg om valg av hvilke klassenumre de forskjellige Humord-emnene bør mappes mot, samt vurdering av relasjonstyper. For at disse erfaringene skal kunne mates inn i mappingverktøyet, gjenstår da en vurdering av graden av sammenfall mellom hvert Humord-emne (inkludert kontekst) og klassenummeret det mappes mot (med sin kontekst i WebDewey). Vi vurderer også i hvilke situasjoner vi ser at andre kilder (som UBO emneregister til Dewey og katalogdata) vil bidra med nyttig kontekst – kontra støy – i mappingverktøyet.

Det intellektuelle bidraget ved datastøttet mapping

Ønsket om å utvikle en metodikk for mapping som et datastøttet intellektuelt arbeid, innebærer en ambisjon om å utvikle et verktøy som gir beslutningsgrunnlag for to typer valg: For det første, hvilke emneord i Humord skal mappes mot hvilke klassenumre i WebDewey? For det andre, hvilken relasjonstype skal den enkelte koblingen betegnes med?

La oss se nærmere på det første spørsmålet, om hvilke elementer som bør kobles sammen: Ideelt sett skal mappingverktøyet komme opp med aktuelle mappingkandidater med synkende relevans. Når dette ikke alltid lar seg gjøre, konfronteres man med støy: For den som foretar det intellektuelle bidraget i mappingarbeidet, vil man i noen tilfeller måtte vurdere mange irrelevante mappingforslag. Samtidig kan fravær av relevante mappingkandidater utgjøre en stor utfordring: Dersom de som foretar mappingene skal kunne fange opp at aktuelle koblinger mangler i lista over mappingkandidater, vil dette forutsette inngående kjennskap til Dewey.

En sentral problemstilling i prosjektet er å utforske potensialet for automatiserte metoder i mappingarbeidet. Hvor går grensen mellom hva det er hensiktsmessig å automatisere, og hvilke operasjoner vil uansett måtte kreve en intellektuell vurdering? Et eksempel på noe som vanskelig lar seg automatisere, er mapping mot bygde numre. Dette vil til gjengjeld være svært nyttig å få gjort, med tanke på at resultatet av mappingen er tenkt å framkomme som Humord-emner i WebDewey, som dermed vil bli beriket med et stort antall ferdigbygde numre innenfor Humords domene.

Når det gjelder valget av relasjonstyper, vil dette være en intellektuell operasjon. Bidraget fra mappingverktøyet vil være å vise konteksten for kandidatene i kilde- og målvokabularet. I den forbindelse jobber vi med å avklare hvilke elementer i hvert av vokabularene som gir det beste beslutningsgrunnlaget – for mye input vil bare gi mental overbelastning.

Ifølge ISO-standarden vil ekvivalensrelasjonen være den mest brukte i mappingarbeid. Dette vil antakeligvis gjelde i større grad ved mapping mellom to tesauri – ikke mellom en tesaurus og et klassifikasjonsskjema. I vårt prosjekt ser det foreløpig ut til at exactMatch sjelden er aktuelt, med unntak av eksempler som topptermen «samfunnsvitenskap» i Humord som kan

mappes mot klasse 300 i Dewey. Klassebetegnelsene representerer ofte en klynge av begreper, og dermed vil en exactMatch være uaktuell. Derimot står vi ofte overfor problemstillingen om man skal betegne en relasjon som closeMatch eller broadMatch. Valget av relasjonstype vil få konsekvenser for senere anvendelser i gjenfinning, for eksempel hvis man ønsker å utnytte mappingene i automatisert utvidelse av søk. Slike problemstillinger tar vi med oss til neste seminar i EDUG-samarbeidet. Mappingseminaret i april har som målsetting å etablere felles retningslinjer for bestepraksis for mapping, fra et sluttbrukerperspektiv.

I drøftingen over har vi sett eksempler på ekvivalensrelasjoner (exactMatch og closeMatch) og hierarkiske relasjoner (broadMatch og narrowMatch) – men når vil det være aktuelt med assosiative mappingrelasjoner (relatedMatch)? Ifølge standarden kan man etablere en assosiativ relasjon dersom kildebegrepet kan assosieres med målbegrepet på en slik måte at man kan anta at målbegrepet (i vårt tilfelle representert ved et klassenummer) vil være relevant for den som søker etter kildebegrepet (Humord-emneordet). Standarden eksemplifiserer med forholdet mellom «e-læring» og «fjernundervisning». Samtidig sier den at det vil være et uskarpt skille mellom en relatedMatch og en closeMatch, men at man må foreta pragmatiske valg i forhold til blant annet brukerperspektivet.

I testmappingen har vi foreløpig brukt relatedMatch når vi lager mapping mot klassenummeret for relaterte termer (se også-henvisninger til en foretrukken term i Humord). I det tidligere nevnte eksemplet med «samlinger (bibliotek)» som ble mappet mot 026 og 027, vil det for eksempel være aktuelt å vurdere en relatedMatch til 025.21 («samlingsutvikling»). Dette begrepet tilhører en annen begrepskategori (operasjon kontra objekt), men en mapping kan være nyttig for sluttbruker ved søk på «samlinger».

Vi har sett at ISO-standarden legger opp til et skille mellom fem ulike typer mappingrelasjoner, men at det praktiske mappingarbeidet reiser mange problemstillinger i valget av relasjonstype i det enkelte tilfelle. Her gjenstår et arbeid med etablering av felles retningslinjer. Uten konsekvent mappingpraksis vil man ikke ha glede av skillet mellom ulike relasjonstyper i forbindelse med anvendelse i sluttbrukerverktøy. Videre testmapping vil avklare om det kanskje vil være nyttig å operere med et lavere antall relasjonstyper, f.eks. kun closeMatch og narrowMatch.

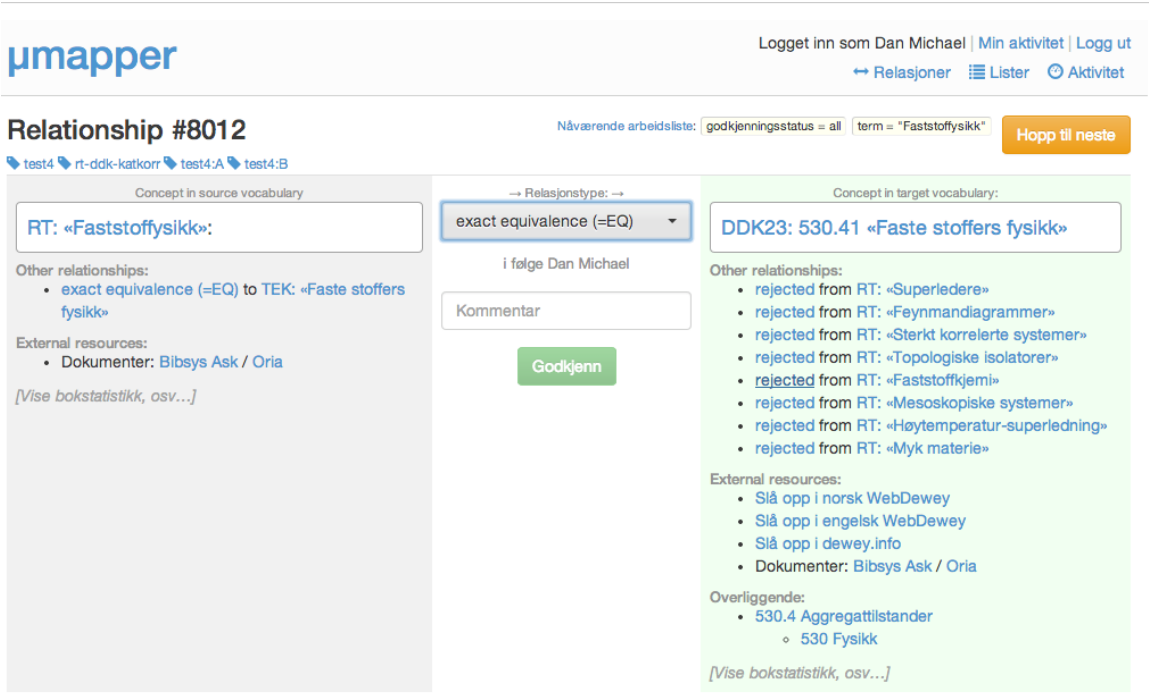
Begge elementene i det intellektuelle bidraget i datastøttet mapping viser seg å være svært utfordrende – både valg av hvilke elementer som skal kobles, og hvordan disse skal kobles. Formålet med mappingverktøyet er å tilby avlastning (ikke *be*lastning) i forbindelse med valg av mappingkandidater og relasjoner. Når vi i mappingarbeidet må foreta pragmatiske valg, er målsettingen å la disse være styrt av sluttbrukerperspektivet: Hva vil gi den beste emnemessige inngangen til informasjonsressursene?

Design av mappingverktøyet ccmapper

Som nevnt i kapittelet om intellektuelle utfordringer ved mapping, ser vi for oss mapping som et datastøttet intellektuelt arbeid, og mappingverktøyet skal gi støtte på to hovedområder:

1. Forslag til mest relevante mappingkandidater for et emneord
2. God oversikt over kontekst for begreper i både kilde- og målvokabular

Prosjektet bygger på erfaringene fra mapping av Realfagstermer som kildevokabular mot den foreløpige norske oversettelsen av Dewey Decimal Classification (DDC) som målvokabular. Her ble prototypen µmapper utviklet, jfr. illustrasjonen i figur 2:



Figur 2: Skjermbilde fra µmapper

Som tidligere nevnt er Realfagstermer en liste av emneord med flat struktur, mens vi i dette prosjektet også har Humord som kildevokabular – en tesaurus med en hierarkisk struktur.

Grunnlaget for automatiske mappingforslag i µmapper var termlikhet mellom Realfagstermer og klassebetegnelser i DDC. Det ble også gjort forsøk med statistisk mapping basert på korrelasjon mellom Realfagstermer og DDC-numre i katalogposter, men datagrunnlaget var for lite. Litteraturbelegg som kontekst ble imidlertid viktig i den intellektuelle vurderingen for å avgjøre en terms betydningsinnhold der det var nødvendig. Vurdering av målbegrepenes plassering i deweyhierarkiet ble også fullstendig overlatt til intellektuell vurdering, men begrensningen til 500 og 600-640-gruppene gjorde vurderingene mindre omfattende enn de ville blitt om hele deweyskjemaet skulle vært inkludert.

Sammenliknet med Realfagstermer har Humord mye mer kontekst som vi må ta hensyn til for å kunne generere gode forslag, i og med at en term i Humord består av emneordet, eventuelle synonymer og hierarkisk kontekst. Dette gjør at vi står ovenfor en del nye utfordringer i dette videreføringsprosjektet.

En deweyklasse representerer på sin side et begrep som er uttrykt ved et klassenummer med en beskrivende klassebetegnelse, hierarkisk kontekst, andre betegnelser (inkl. synonymer), samt noter.

> Mappings to or from a class or category in a monohierarchical scheme should treat the class/category as a pre-coordinated concept whose meaning can be established by inspecting all its superordinate and subordinate classes as well as any scope notes associated with it. Inspection of the caption alone is inadequate.
>
> (International Organization for Standardization, 2013, s.32)

Det ble tidlig klart at det var behov for en mer avansert løsning enn en tradisjonell termsammenligning. For å komme opp med gode mappingforslag er det en stor fordel om mappingverktøyet kan forholde seg til begreper, ikke termer, gjennom å ta hensyn til kontekst for både emneordet og deweyklassen.

ISO 25964-2, kapittel 14, «Techniques for identifying candidate mappings» og kapittel 14.2 «Computer assisted direct matching» beskriver anbefalt fremgangsmåte og behovet for å gi best mulig oversikt over kilde- og målvokabular:

> The candidate mappings identified by the matching processes described should be assembled for review by an expert. For each concept in the source vocabulary, the expert should be able to view the complete record (including scope note, broader and narrower terms).
>
> […]
>
> The viewing interface should make it easy to check the complete context of each concept identified in the target vocabulary. It should also support the expert in selecting the appropriate type of mapping for the candidate(s) he approves.
>
> (s.40)

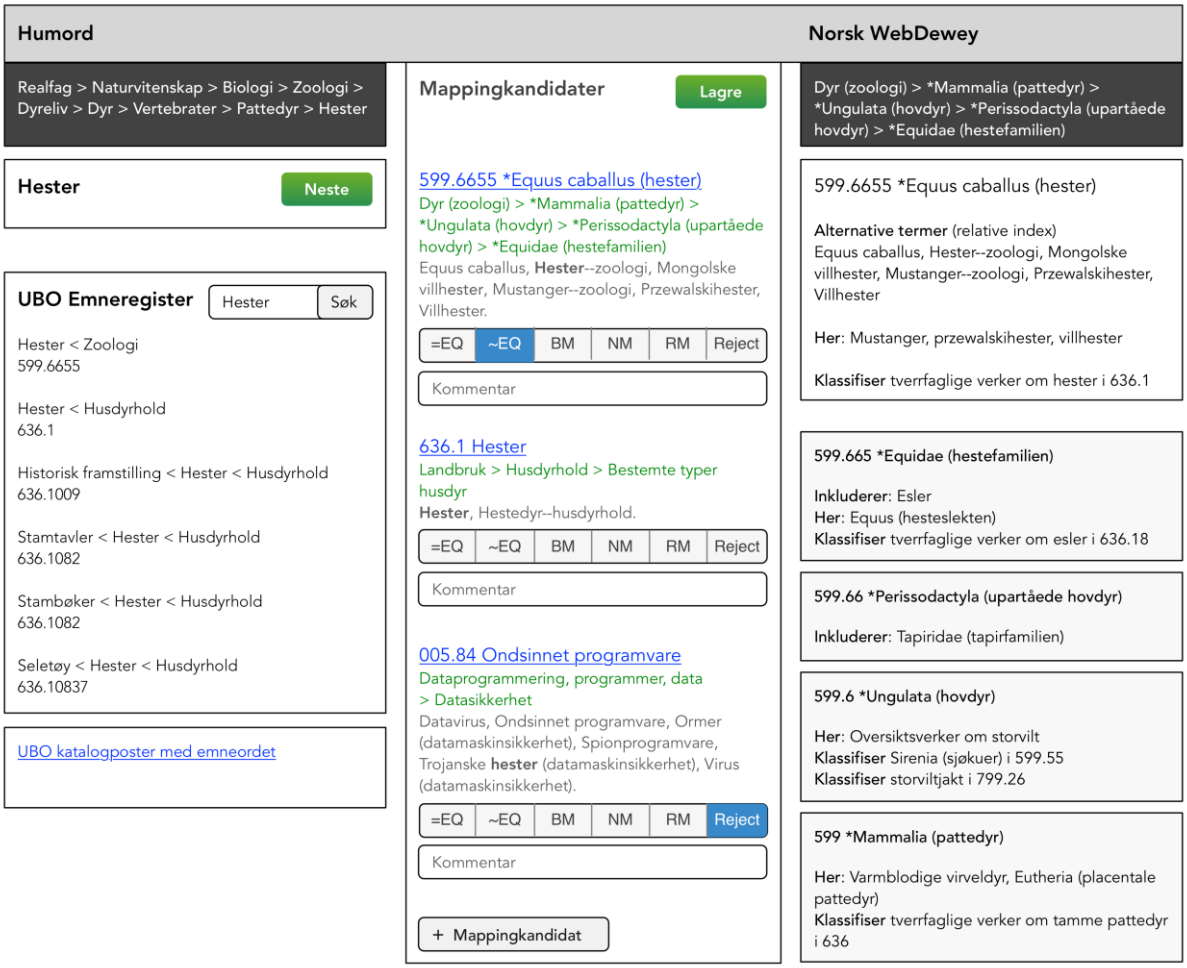
Mappingverktøyet skal altså gi forslag i form av en liste av mappingkandidater som må vurderes av en ekspert. Eksperten bestemmer riktige mappingrelasjoner for ønskede mappingkandidater og forkaster andre mappingkandidater.

For hver mappingkandidat må verktøyet kunne gi eksperten en kompakt og best mulig oversikt over kilde- og målbegrepets betydningsinnhold og kontekst for å støtte opp under vurderingen av kandidatrelasjonen.

Vi har valgt brukergrensesnittmetaforen «dashbord». Figur 3 på neste side illustrerer hvordan dette er tenkt per februar 2015. Godt interaksjonsdesign vil være en viktig suksessfaktor.

Brukergruppen for verktøyet er bibliotekarer. Vi kan dermed tillate oss å lage et verktøy med en mer kompleks oppbygning av brukergrensesnittet enn for en vanlig webapplikasjon. Personene som gjør arbeidet er eksperter innen fagområdet og vil få opplæring.

I tillegg til kilde- og målvokabular utnytter vi andre datakilder som enten kan gi bedre kontekst for begrepene som mappes, eller gi statistisk informasjon om faktisk bruk av kombinasjoner av emneord og deweyklasser i bibliotekkatalogen. Universitetsbiblioteket i Oslo har bl.a. et Dewey emneregister som vi utnytter i tillegg til uttrekk av katalogdata.



Figur 3: Interaksjonsdesign for prototypen ccmapper per februar 2015.

Applikasjonsnavnet ccmapper står for concept context mapper. Vi planlegger å få ferdig en prototype i løpet av våren 2015.

Under beskrives og drøftes problemstillinger knyttet til realisering av mappingverktøyet.

Datakilder, SKOS og lenkede data

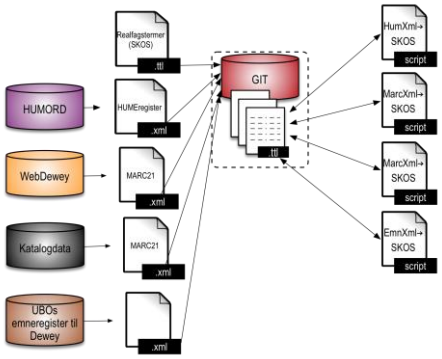
Som del av mappingforberedelsene har vi samlet inn og konvertert data fra de ulike datakildene vi planlegger å bruke:

Tesaurusen Humord

Norsk WebDewey

UBO emneregister til Dewey

Katalogdata for UBO



Figur 4: Datakilder

Alle datakildene har i løpet av prosjektet blitt konvertert til SKOS og lagt i et GIT-arkiv[5] for versjonshåndtering.

Det meste er publisert åpent med CC0 1.0-lisens[6]. Kildefiler og konverterte filer for norsk WebDewey er ikke publisert åpent på grunn av opphavsrettslige forhold.

Humord, UBO emneregister til Dewey og Realfagstermer er publisert som åpne lenkede data på data.ub.uio.no.

Vi regner med å gjøre mindre endringer i kodingen fortløpende ut fra erfaringer i dette prosjektet. Vi må også se an samarbeidsprosjektet «BIBSYS Biblioteksbase i semantisk web» rundt publisering av katalogen som åpne lenkede data, som har fått støtte fra NB for 2015. Dette er et samarbeid mellom BIBSYS og UiO, UiB, NTNU og UiT Norges arktiske universitet bl.a. mot én felles RDF-representasjon, istedenfor flere lokale varianter.

Vi må også se an hvor og hvordan dataene skal publiseres på sikt. Hvis Nasjonalbiblioteket blir eierorganisasjon for en nasjonal tesaurus bør dette publiseres innenfor NBs domener, f.eks. data.nb.no, eller eventuelt under data.norge.no.

SKOS (Simple Knowledge Organization System)

Siden vi arbeider med data fra ulike kunnskapsorganisasjonssystemer, passer W3C-standarden SKOS godt som felles datamodell. SKOS er et RDF-vokabular, og koding av datakildene som lenkede data gir oss svært gode muligheter for å søke i og koble data på tvers av ulike systemer.

SKOS er laget for å kunne modellere vanlige, men relativt enkle og semi-formelle kunnskapsorganisasjonssystemer som tesauri, taksonomier, klassifikasjonssystemer, folksonomier og kontrollerte vokabularer.

Standarden inneholder også relasjoner for mapping.

> SKOS occupies a position between the exploitation and analysis of unstructured information, the informal and socially-mediated organization of information on a large scale, and the formal representation of knowledge.[7]

Semantikken er ikke formelt definert i SKOS. For å være generell, definerer SKOS med vilje ikke i detalj hva mappingrelasjonene betyr. Man bruker her begrepet «Emphasis on minimal ontological commitment».

SKOS er ment å være et bindeledd mellom mer uformelle informasjonsstrukturer og ontologier som gjerne er basert på logikk og aksiomer.

Den overordnede SKOS-modellen for å modellere kunnskapsorganisasjonssystemer er relativt enkel (ikke inkludert mapping-relasjonene):

[5] https://utv.uio.no/stash/projects/UB/repos/datakilder/
[6] http://creativecommons.org/publicdomain/zero/1.0/
[7] SKOS W3C Recommendation, kap. 1.1, Background and Motivation.http://www.w3.org/TR/skos-reference/#L879

- Modellering av entiteter
  - Concept skos:Concept
  - Label (termer for å referere til concept):
    - Preferred Label skos:prefLabel
    - Alternative Label skos:altLabel
    - Hiddel Label skos:hiddenLabel

- Modellering av relasjoner
  - Broader/Narrowerskos:broader, skos:narrower
  - Assosiativ relasjon skos:related

- Dokumentasjon, dvs. ulike typer noter
  - skos:scopeNote, skos:definition, skos:example, skos:historyNote, skos:editorialNote, skos:changeNote

Det er fem mappingrelasjoner i SKOS, som diskutert tidligere i rapporten og illustrert i interaksjonsdesignet for prototypen over.

ISO 25964-2	Forkortelse	SKOS
Exact equivalence	=EQ	skos:exactMatch (symmetric, transitive)
Inexact equivalence	~EQ	skos:closeMatch (symmetric)
Broader mapping	BM	skos:broadMatch (inverse of narrowMatch)
Narrower Mapping	NM	skos:narrowMatch (inverse of broadMatch)
Related Mapping	RM	skos:relatedMatch (symmetric)

Datakilder

Som del av mappingforberedelsene har vi samlet inn og konvertert data fra de ulike datakildene vi planlegger å bruke.

Realfagstermer

Realfagstermer forelå allerede ved prosjektstart som RDF/SKOS med CC0-lisens. I løpet av høsten 2014 har vokabularet også fått et driftsopplegg på data.ub.uio.no slik at de åpne dataene alltid er oppdaterte.

Humord og UBO emneregister til Dewey

Humord og UBO emneregister til Dewey forelå ikke som åpne data ved prosjektstart. Begge vedlikeholdes i EMNE-modulen i BIBSYS som har støtte for XML-eksport. Vi har konvertert begge til RDF/SKOS og publisert dataene med CC0-lisens på data.ub.uio.no. Et eksempelbegrep fra Humord i SKOS (og Turtle-serialisering) ser slik ut:

```
<http://data.ub.uio.no/humord/c05316> a skos:Concept ;
   dct:identifier "HUME05316" ;
   dct:modified "1994-03-21"^^xsd:date ;
   skos:altLabel "Bildende kunst"@nb,
       "Billedkunst"@nb,
       "Visuell kunst"@nb ;
   skos:broader <http://data.ub.uio.no/humord/c05183> ;
   skos:definition "Bildekunst omfatter tradisjonelt visuell kunst: malerkunst,
tegnekunst, grafisk kunst, bildehoggerkunst og bildevev. Her også nye medier som
fotokunst, videokunst mm <UBB>"@nb ;
   skos:inScheme <http://data.ub.uio.no/humord/> ;
   skos:prefLabel "Bildekunst"@nb .
```

Med konverteringen har vi gått fra en termbasert modell i ISO 2788-tradisjon til en begrepsbasert modell der hver term har en foretrukket term (skos:prefLabel) og null eller flere alternative termer (skos:altLabel). Se-henvisninger konverteres til alternative termer ("Sykkelstier" SE "Sykkelveier" blir til skos:prefLabel "Sykkelveier"; skos:altLabel "Sykkelstier").

Den eneste virkelige utfordringen for modellskiftet er generelle se-henvisninger: henvisninger fra én term til to begreper (for eksempel henvisningen «Buddhistisk filosofi, SE Buddhisme * Filosofi» i Humord). Hvis disse skal uttrykkes i RDF må det gjøres på en måte som er fjernt fra SKOS-modellen (ISO 25964-utvidelsen til SKOS skisserer én slik mulighet). Generelle se-henvisninger er et indre anliggende i Humord-tesaurusen og ville uansett ikke være aktuelle å mappe mot Dewey.

Andre utfordringer for konvertering til RDF/SKOS inkluderer fasettindikatorer og knutetermer, men for mappingformål er ingen av disse klassene særlig interessante fordi de representerer tesaurustekniske elementer som ikke brukes til indeksering. Vi har valgt å inkludere de i konverteringen for visningsformål, men har gitt dem egne klasser så de enkelt kan skilles fra vanlige emneord.

Norsk WebDewey

Norsk WebDewey er ikke ferdig, men fra NB har vi fått tilgang til å eksportere status quo fra oversettelsesverktøyet som MARC21XML til internt bruk. Disse har blitt konvertert til RDF/SKOS med noen lokale utvidelser. Utover modellen til dewey.info har vi inkludert registertermer (som skos:altLabel) og noter. Fire notetyper er spesielt interessante fordi de inneholder referanser til termer tydelig adskilt i egne delfelt:

Notetype	Eksempel i MARC21 Classification	Eksempel konvertert til RDF
"Andre betegnelser" (Variant-names, ess=nvn)	680$iAndre betegnelser:$tVaskulæare kryptogamer$i,$tvaskulære planter uten frø$9ess=nvn	wd:variantName "Vaskulæare kryptogamer"@nb, "Vaskulære planter uten frø"@nb
”Her” (Class-here, ess=nch)	680$iHer:$tAssosiative algebraer$i,$tikke-kommutative algebraer$t,ikke-kommutative ringer$9ess=nch	wd:classHere "Assosiative algebraer"@nb, "Ikke-kommutative algebraer"@nb, "Ikke-kommutative ringer"@nb
"Inkluderer"-noter (Including, ess=nin)	680$iInkluderer:$tKorrekturlesing$9ess=nin	wd:including "Korrekturlesing"@nb
"Tidligere klassebetegnelse" (Former heading, ess=nph)	680$iTidligere klassebetegnelse:$tKidneybønner$9ess=nph	wd:formerHeading "Kidneybønner"@nb

Mer beskrivende noter som forklarende noter (definition note, ess=ndf) og omfangsnoter (scope notes, ess=nsc) er også konvertert, men vi er mer usikre på om disse kan utnyttes i mappingsammenheng.

Mappingskjema og beskrivelse finnes på https://github.com/scriptotek/mc2skos. Selve dataene har vi ikke tillatelse til å publisere.

Indekserings- og klassifikasjonsdata fra bibliografiske katalogposter

For hvert par av begreper fra ulike vokabularer (som f.eks. Humord og DDC) kan vi finne en statistisk assosiasjonsgrad basert på hvor ofte paret opptrer sammen i forhold til hvor ofte hvert medlem opptrer alene i bibliografiske katalogposter. Fra et tidligere prosjekt har vi en dump med data til og med april 2014, og vi kommer til å be BIBSYS om en ny dump når det nærmer seg oppstart av mappingen.

På katalogpostene finnes det deweynumre fra mange forskjellige utgaver. Det hadde vært ideelt å kunne avgrense til én bestemt utgave som DDC-23, men vi har bare ~ 50 000 poster som har både DDC-23 og Humord og ~ 7 000 poster som har både DDC-23 og Realfagstermer. I forhold til størrelsene på de involverte vokabularene er dette for lave tall.

For å øke størrelsen på datagrunnlaget planlegger vi å heller gjøre en avgrensning i tid. Avgrenser vi f.eks. til poster f.o.m. år 2000 gir det oss ~ 200 000 poster med Humord + DDC og ~ 15 000 poster med Realfagstermer + DDC. For Realfagstermer er datagrunnlaget fremdeles for lite, men for Humord vil det bli interessant å se hvorvidt statistisk mapping kan gi gode forslag.

Automatisk generert liste med mappingkandidater

Vi har sett nærmere på flere verktøy som er i bruk for å løse lignende problemstillinger[8]. Det viser seg å være svært mange beslektede begreper, tilnærminger og fagområder som er delvis overlappende.

Innenfor lenkede data er begrepene «Equivalence Mining», «Equivalence Matching» og «link discovery» etablert. Her ser verktøyene Silk («A Link Discovery Framework for the Web of Data») og LIMES («Link Discovery Framework for Metric Spaces») ut til å være mest brukt.

Hvis det er snakk om å lenke datasett som ikke bruker felles identifikatorer, spesielt kobling av databaseposter i ulike systemer er følgende begreper i bruk: «Record linkage», «Entity resolution», «Name resolution», «Identity resolution», «Deduplication» og «Merge/purge».

Felles for verktøyene og tilnærmingene vi har sett på er at de i stor grad baserer seg på sammenligning av termlikhet.

Andre beslektede tilnærminger og fagområder er bl.a. «Ontology Mapping», «Ontology Alignment», «Ontology Matching», «Semantic Matching» og «Semantic Mapping». *Ontology Matching* (Euzenat & Shvaiko, 2013) gir en systematisk oversikt over kjente teknikker og strategier.

Vi ønsker en datagenerert liste med forslag til mappinger som på en fleksibel måte tar hensyn til kontekst for både termen og klassen. Vi ønsker også enkelt å kunne endre vekting for ulike former for kontekst, og listen bør være sortert etter relevans.

Vi har derfor valgt å bruke vektorromsmodellen, som gir oss mulighet til å ta hensyn til alle former for kontekst og uavhengig vekting av disse på en enkel og standardisert måte.

For hvert begrep i kilde- og målvokabulane planlegger vi å generere et syntetisk tekstdokument som representerer begrepet. Dokumentet vil inneholde tekstlig representasjon av alt som utgjør innhold og kontekst for begrepet og alle elementer kan vektes opp ved å gjentas et antall ganger.

Gjennom å bruke vektorromsmodellen (Salton, Wong, & Yang, 1975) som er allestedsnærværende i moderne indeksering og søk, kan vi benytte oss av standard søketeknologi som Apache Lucene[9] for å indeksere og søke i dokumentene.

For hvert emneord kan vi bruke den tilhørende termvektoren som søkevektor og rangere termvektorene fra deweydokumentene etter cosinus av vinkelavstanden til termvektoren for emneordet (dvs. standard vektorbasert søk).

Vi bruker videre TF-IDF (Term frequency-Inverse document frequency) til å vekte termene. TF-IDF som gir høy vekting til ord som forekommer sjelden i dokumentsamlingen. Enkelt forklart vil dette si at ord som bidrar til å skille de ulike begrepene fra hverandre vektes

8 http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining
9 http://lucene.apache.org

høyere enn ord som forekommer ofte (f.eks. i mange ulike deweyklasser) og ikke bidrar til å skille begrepene fra hverandre.

Vi ser at vi må normalisere termene i indeksen, dvs. redusere dem til roten, for å klare å få treff fra emneord til riktig deweyklasse. Vi ønsker derfor å bruke en nyutviklet lemmatiseringsmodul for Lucene/solr som bruker Språkbankens ordlister.

I tillegg ser vi at det i endel tilfeller vil være nødvendig å få til splitting av sammensatte ord for å få riktige treff. Artikkelen *Monolingual document retrieval for European languages* (Hollink, Kamps, Monz, & De Rijke, 2004) viser at ordsplitting («compund splitting») for svensk gir hele 25 % bedre presisjon («mean average precision») i forhold til standard normalisering av ordstammen.

Vi er imidlertid på dette stadiet usikre på om det er realistisk å få på plass ordsplitting i prosjektet.

En kort oppsummering av metoden:

- Bruk term og tilgjengelig kontekst til å lage et syntetisk tekstdokument som er en representasjon av begrepet som skal mappes
    - Filtrer ut stoppord
    - De ulike bestanddelene av det syntetiske dokumentet må vektes. (Termen må f.eks. vektes høyere enn overordnede termer)
    - Man kan inkludere kontekst fra relaterte systemer i det syntetiske dokumentet, f.eks. fra UBOs emneregister for Dewey og fra forekomster av emneord og deweyklasser i katalogen

- Indekser tekstdokumentene med Lucene og bruk tf–idf (term frequency–inverse document frequency)[10]
    - Eventuell splitting av sammensatte ord (legg oppsplittede termer til indeksen)
    - Stemming/lemmatisering av ord for å få enhetlige ordstammer
- Bruk dokumentvektor for det aktuelle emneordet som søkevektor ved søk mot dokumentindeksen for deweyklasser
- Sorter trefflisten etter cosinus-likhet for vinkelen mellom vektorene

Bruk av emneregister og katalogdata som kontekst

Data fra UBO emneregister til Dewey brukes til å gi kontekst i brukergrensesnittet, ref. interaksjonsdesignfigur over.

UBO emneregister er bygget opp etter kjederegistermetoden og setter termen inn i en faglig sammenheng. Når vi finner sammenfall mellom en term i Humord og tilsvarende i UBO-registeret, vil koblingen mellom emneord og klassenummer gi en sterk kandidat for mapping. Vi vil derfor inkludere klassenummeret i det syntetiske dokumentet.

Tilsvarende kan analyse av sammenfall av emneord og deweyklassifikasjon av poster i katalogen utnyttes. Dette kalles ofte for statistisk mapping. I vårt tilfelle har det sine begrensinger da tilnærmingen vår bl.a. må dekke mapping der ikke alle postene har deweyklassifisering.

10 http://en.wikipedia.org/wiki/Tf–idf

Vi har imidlertid også mange forekomster av poster med flere emneord og flere deweyklasser. I disse tilfellene vil koblingen mellom emneord og deweyklasse være usikker.

Vi ser også at emner fra BIBBI, Biblioteksentralens bibliografiske database, ville kunne hjelpe oss. Basen inneholder ca. 220 000 katalogposter som er blitt tildelt både emneord og deweynumre (det siste riktignok i tråd med forkortete, norske tabeller – DDK5). For delmengden av emneord i BIBBI som er sammenfallende med Humord, ser vi at vi vil kunne få god hjelp til å foreslå de mest relevante mappingkandidatene for et emneord.

Det er fortsatt et åpent spørsmål om man ønsker en generell eller en mer kontekst-avhengig mapping i koblingen mellom Humord og WebDewey. Utfordringen ved bruk av både UBO emneregister og statistisk mapping som kontekst i de syntetiske dokumentene for Humord-termene, er at vi kan få sterke koblinger mot klasser som i forhold til kontekst i Humord ikke gir riktig mapping. Hvis vi ønsker en generell mapping mot alle faglige perspektiver i DDC vil denne konteksten være nyttig, mens det vil introdusere støy hvis vi kun ønsker mapping basert på Humord-kontekst for emneordet. Dette ble illustrert med eksempelet om «hester» i kapittelet «Intellektuelle utfordringer i forbindelse med mapping».

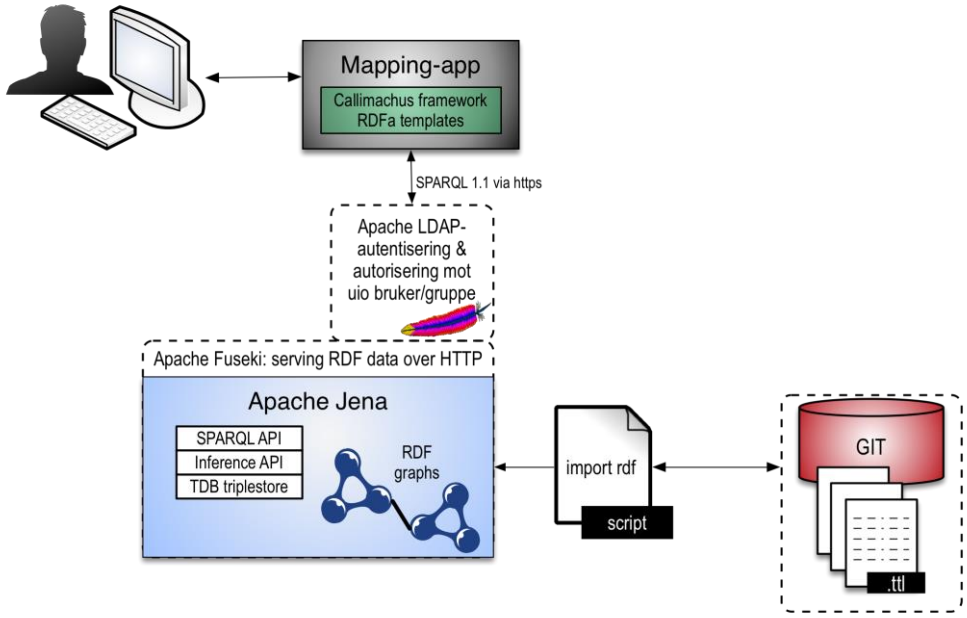
Dette viser imidlertid at metoden vil kunne brukes i begge tilfeller, men at man må være klar over hvordan bruk av emneregistre og statistikk fra katalogen vil påvirke resultatet.

Bruk av emneregisteret fra BIBBI vil sannsynligvis føre til en enda mer generell mapping som vil reflektere dokumentsamlingen i et folkebibliotek.

Applikasjonsarkitektur for ccmapper

Prototypen av ccmapper skal som nevnt være basert på SKOS og SPARQL. I utvikling av prototypen bruker vi rammeverket Callimachus (callimachusproject.org), som er basert på RDFa, SPARQL, XHTML5, CSS3 og JavaScript.

Det er foreløpig for tidlig å konkludere med om dette rammeverket også vil være egnet til den endelige versjonen av ccmapper som skal settes i drift.



Figur 5: Overordnet systemskisse for ccmapper i produksjon

Avslutning

I den foreliggende rapporten redegjør vi for aktiviteter som er sluttført eller under arbeid per utgangen av februar 2015. Det er viktig å peke på at en del av de aktivitetene og problemstillingene vi har reist ikke vil avsluttes innen utgangen av inneværende prosjekt, men vil bli ført videre i løpet av inneværende år.

Universitetsbiblioteket i Oslo sendte høsten 2014 en ny prosjektsøknad til Nasjonalbiblioteket om midler til å planlegge og gjennomføre mapping mot DDC i norsk oversettelse (norsk WebDewey). Vi har fått innvilget budsjettmidler. Dette har gitt oss en unik mulighet til å videreføre metodikkprosjektet inn i det nye prosjektet *Mapping mot norsk WebDewey.*

Mappingarbeidet, som vil omfatte vokabularene Humord og Realfagstermer, vil utføres i tråd med ISO 25964-2 (International Organization for Standardization, 2013). Som beskrevet tidligere inneholder denne ISO-standarden generelle retningslinjer, ikke spesifikke retningslinjer for mapping mot et klassifikasjonssystem som DDC. Det er derfor behov for å utvikle teoretisk forståelse og praktiske fortolkninger som er tilpasset våre vokabular, både kildevokabularene Humord og Realfagstermer og målvokabularet DDC.

Under det årlige EDUG Mapping Working Group Meeting i Reykjavik i 2014, ble det klart for oss at det er en generell interesse for å gjøre denne type avklaringer. Hva innebærer det i teori og praksis å mappe et emneordsvokabular inn mot DDC? Det hersket ikke minst en del usikkerhet rundt hvordan ulike valg av mappingrelasjoner vil kunne påvirke framtidig funksjonalitet i sluttbrukerverktøy.

Vi planlegger derfor i samarbeid med EDUG et mappingseminar i forbindelse med EDUGs årlige møte i 2015. Seminaret vil gå over halvannen dag (15. og 16. april i Napoli) og har fått den foreløpige tittelen *Mapping to Dewey: Recommendations for Best Practice.* Målgruppen for seminaret er ulike miljøer som har erfaring med, eller interesse for, mapping av ulike vokabular til DDC. Vi ønsker at utfallet av seminaret blir et sett med anbefalinger om bestepraksis på området med utgangspunkt i ISO-standarden om mapping. Til seminaret har vi invitert medlemmer av arbeidsgruppa bak mappingstandarden for en presentasjon av standarden og teoretiske forhold som må tas i betraktning når det mappes mot et klassifikasjonsskjema som DDC. Vi har også invitert spesialister på SKOS og Dewey, miljøer som har mappet mot DDC og miljøer som har erfaring med utvikling av mappingverktøy. Det er avsatt tid til diskusjoner bade i plenum og gruppevis hvor også brukerperspektivet trekkes inn. Diskusjonene vil danne grunnlag for utarbeidelse av generelle retningslinjer.

Retningslinjene som kommer ut av seminaret i Napoli vil få konsekvenser for hvordan UBO i neste omgang tenker den faktiske mappingen utført. Vi ser for oss at det i etterkant av seminaret vil være behov for å gjøre justeringer på dette området.

Under seminaret i Napoli planlegger vi også et møte med Pansoft som har utviklet oversetterprogram og distribusjonsserver for de ulike oversettelsene av Dewey Decimal Classification. Pansoft har uttrykt interesse for en samhandling og mulig innlemmelse av prosjektets prototype i Pansofts portefølje. Dette er i tråd med UBOs mål om deling av kompetanse. Utvikling av en produksjonsløsning har hele tiden vært definert utenfor vårt prosjekt.

Algoritmene som skal foreslå kandidater i mappingverktøyet, kommer til å bli utarbeidet med utgangspunkt i de erfaringene vi skaffer oss gjennom testmappingen. Verktøyet kommer videre til å bli justert i tråd med anbefalingene for mapping som kommer ut av Napoli-

seminaret. Deretter vil vi starte med den faktiske mappingen mellom Humord og Realfagstermer til DDC**.** Vi vil jobbe tett med Nasjonalbibliotekets deweyredaksjon om dette. Vi mener at arbeidet vårt vil ha overføringsverdi til andre institusjoner i Norge og internasjonalt som ønsker å utføre tilsvarende mappinger mot DDC, og vil legge til rette for å dele våre erfaringer med andre miljøer.

Referanser

Euzenat, J., & Shvaiko, P. (2013). *Ontology matching* (2. utg.), Berlin: Springer.

Hollink, V. Kamps, J., Monz, C, & De Rijke, M. (2004, januar). Monolingual document retrieval for European languages. *Information Retrieval*, *7*(1-2), 33-52. Doi:10.1023/B:INRT.0000009439.19151.4c

International Organization for Standardization. (2009). *Information and documentation: Thesauri and interoperability with other vocabularies: Part 1: Thesauri for information retrieval (ISO 25964-1: 2011).* Geneve: International Organization for Standardization.

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

Knutsen, U., & Gulbrandsen, A.D. (2014). På randen av mapping. *Bibliotheca Nova,* (4), 36-46.

Kuldvere, V., Lundevall, M., Hegna, K., Konestabo, H. S., Låberg, K. T. , Flatby, E. S., & Greenall, R. (2013, 7. mai). *Realfagstermer og TEKORD: RDF som plattform for sammenlikning og sammenføyning av emnesystemer?: Rapport*. Hentet fra 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, 1. juli). *Felles terminologi for klassifikasjon med Dewey: Rapport*. Hentet fra http://urn.nb.no/URN:NBN:no-44610

Salton, G., Wong, A., & Yang, C. S. (1975, november). A vector space model for automatic indexing. *Communications of the ACM,18*(11), 613–620.

Vedlegg: Regnskap

	UiO Prosjektleder:KNUTSEN, UNNI	Eksternt finansiert OH%:

		Regnskap hittil i år
	Eksterne inntekter og kostnader	
	Eksterne inntekter	-1 350 000
	Totale eksterne inntekter	-1 350 000
	Personalkostnader	508 095
	Direkte driftskostnader	47 178
	Totale eksternt finansierte kostnader	555 273
	Saldo/Resultat på prosjekt	-794 727
	Internfinansiering av ressurser	
	Egenandel	-593 411
	Frikjøp internt finansiert	518 799
	Overhead internt finansiert	409 456
	Sum interne inntekter og kostnader	334 844
	Akkumulert saldo	-459 883

	UiO Prosjektleder:KNUTSEN, UNNI	Eksternt finansiert OH%:

		Regnskap hittil i år
	Overført fra ifjor	-459 883
	Eksterne inntekter og kostnader	
	Eksterne inntekter	0
	Totale eksterne inntekter	0
	Personalkostnader	182 152
	Direkte driftskostnader	4 401
	Totale eksternt finansierte kostnader	186 554
	Saldo/Resultat på prosjekt	-273 329
	Internfinansiering av ressurser	
	Egenandel	-143 513
	Frikjøp internt finansiert	70 976
	Overhead internt finansiert	101 251
	Sum interne inntekter og kostnader	28 715
	Akkumulert saldo	-244 614