Realfagstermer og TEKORD

RDF som plattform for sammenlikning og sammenføyning av emnesystemer?

Rapport

Viola Kuldvere, Mari Lundevall, Knut Hegna, Heidi Sjursen Konestabo, Kyrre Traavik Låberg, Ellen Samdahl Flatby, Rurik Greenall

07.05.2013

Innhold

1. Innledning

1.1 Bakgrunn

Prosjektet *Realfagstermer og TEKORD* har hatt som mål å undersøke om RDF (Resource Description Framework)[1] kan fungere som plattform for sammenlikning og sammenføyning av disse to emnesystemene som dekker realfag, naturvitenskap og teknologi.

Prosjektet ble tildelt samarbeids- og utviklingsmidler fra Nasjonalbiblioteket for 2012 (NOK 200 000). Initiativtager og søker om prosjektmidler er Realfagsbiblioteket ved Universitetsbiblioteket i Oslo (UBO). Prosjektet har samarbeidet med Ellen Samdahl Flatby og Rurik Greenall (NTNU Universitetsbiblioteket (NTNU UB)).

Realfagsbiblioteket ved UBO har siden 2009 arbeidet med å utvikle et kontrollert vokabular (Realfagstermer) for matematisk-naturvitenskapelige fag. Fra høsten 2009 til våren 2011 deltok alle avdelinger ved daværende Matematisk-naturvitenskapelig fakultetsbibliotek (nå Realfagsbiblioteket) i arbeidet med overgang fra frie nøkkelord til kontrollerte emneord.

Den økte oppmerksomheten rundt kontrollerte emneord ved Realfagsbiblioteket hadde sitt utspring bl.a. i rapporten *Bibliografisk og emnemessig beskrivelse av UBOs samlinger* [2]. Rapporten peker på den store variasjonen i emneord og klassifikasjon ved Universitetsbiblioteket i Oslo spesielt - og i Norge generelt når det gjelder universitets-og høgskolesektoren. Det brukes mange ulike klassifikasjonssystemer, og emnebeskrivelsen er svært uensartet og fragmentert i ulike kontrollerte vokabularer og fri emneordspraksis (ukontrollerte termer). Dette gir brukerne et dårlig tilbud når det gjelder emnegjenfinning av litteratur - både når det gjelder å få en samlet oversikt over litteraturtilfanget innenfor et emne og når man ønsker presise treff på spesifikke emner.

Samordning og standardisering av emnesystemer og -praksis er nødvendig for at systemene skal kunne ”kommunisere” med andre systemer nasjonalt og internasjonalt, og slik også lette gjenbruk av registreringsdata og forbedre gjenfinningskvaliteten for informasjonssøkeren.

I biblioteksektoren i Norge har man de siste årene blitt mer oppmerksom på fordeler som standardisering og samordning av data vil kunne gi i form av samvirke mellom ulike datamengder og kobling eller mapping av data for å kunne utvikle bedre tjenester for brukerne. RDF og «linked data» [3] nevnes nå hyppig i Norge som et metadataformat som vil være nyttig for kobling av bibliotekdata (for eksempel emneord, og for kobling mot eksterne ressurser).

Det er tidkrevende å vedlikeholde flere systemer, og med tanke på at de fleste kontrollerte emnesystemene i Norge dekker beslektete faglige områder og inneholder struktur som er relativt lik (synonymkontroll, se-henvisninger og hierarki), vil en samordning gi gevinster på ressurssiden. RDF kan muliggjøre samordning av slike data uten at de ulike systemene mister sine særpreg. En viktig fordel her er den enorme veksten dette gir i spennet av informasjon som vil være tilgjengelig for søkesystemene og videre for brukeren.

I internasjonal sammenheng er arbeidet med samordning (blant annet tverrspråklig samordningsarbeid), standardisering, overganger mellom ulike emne- og klassifikasjonssystemer, samt utprøving av semantisk

web-teknologier som linked data kommet et godt stykke lenger enn i Norge (for eksempel FinnONTO - National Semantic Web Ontology Project in Finland [4]). På bakgrunn av det som er gjort internasjonalt, ønsker prosjektet å gi et bidrag til et langsiktig mål om å utvikle en samlet norskspråklig kontrollert emneordsbasert fagterminologi for store deler av kunnskapsuniverset i tråd med gjeldende språkpolitiske føringer. Universitets- og høyskolelovens §1-7 sier: "Universiteter og høyskoler har ansvar for vedlikehold og videreutvikling av norsk fagspråk" [5]. Ved å legge til rette for overganger mellom systemer ved hjelp av mulighetene som ligger i linked data og semantisk web vil deling og gjenbruk av data på tvers av systemer nasjonalt og internasjonalt kunne bli mulig, og informasjonssøkeren vil oppnå effektiv flerspråklig gjenfinning, emnebaserte navigeringsmuligheter og viderelenking til eksterne ressurser. Slik vil gjenfinningskvaliteten for både norske og fremmedspråklige informasjonssøkere ivaretas.

1.2 Prosjektbeskrivelse

Mål

Prosjektets hovedmål har vært å undersøke om RDF kan brukes som plattform for å sammenlikne de kontrollerte emneordene for realfag (Realfagstermer (UBO)) og termene innen teknikk og naturvitenskap (TEKORD (NTNU UB)). Gjennom denne sammenlikningen ønsket vi å se om det er grunnlag for gjensidig berikelse eller utnyttelse, og om vokabularene eventuelt kan slås sammen til ett felles vokabular. Vi har vært interesserte i å høste erfaringer fra prosessen med tanke på om metoden senere kan brukes til å sammenlikne og eventuelt sammenkoble eller sambruke, andre norske emnesystemer.

Prosjektet skulle også se på muligheter for å flette inn andres (eksterne) data som eksisterer som linked open data.

Utover dette har vi ønsket å oppnå samarbeid og idéutveksling rundt mulig framtidig bruk av norske emneord som linked data, og få erfaringer med emneordsarbeid som kan brukes i et nytt biblioteksystem. På sikt vil et slikt samarbeid og en eventuell samordning gi gevinster i brukbarhet for bibliotekansatte og sluttbrukere, og gevinster knyttet til ressursbruk for vedlikehold og videreutvikling av systemene.

Problemstillinger

Vokabularene TEKORD og Realfagstermer har både fellestrekk og ulikheter. Begge vokabularer har ca. 15 000 termer. TEKORD har emneord med en faglig overvekt på teknikk i en hierarkisk struktur med overordnede og underordnede termer (OT og UT), samt registertermer knyttet til UDK, mens Realfagstermer har emner først og fremst innen de naturvitenskapelige fagene, i en flat struktur. Begge vokabularer har synonymer (se-henvisninger) og assosiative relasjoner (se også-henvisninger) til de foretrukne termene. Realfagstermer har i tillegg noe latin og engelsk knyttet til en del termer. Matematiske termer er i stor grad koblet til Mathematics Subject Classification (MSC)[6] og Dewey Decimal Classification (DDC) [7]. I tillegg har Realfagstermer i sitt websøk [8] viderekobling med oppslag i Store Norske Leksikon, Wikipedia og Periodesystemet [9].

Begge vokabularer kan dermed i teorien berikes gjensidig ved å utnytte likheter og ulikheter. Prosjektet kom fram til følgende problemstillinger som utgangspunkt for en sammenlikning med tanke på berikelse/deling:

- Hvor er det overlapp i terminologien til TEKORD og Realfagstermer? Det vil si, hvilke begreper finnes i begge vokabularer (om enn beskrevet med ulike ord)?
- Har TEKORD og Realfagstermer ulike synonymer og se også-henvisninger som kan deles?
- Kan TEKORD få nytte av Realfagstermers engelske termer?
- Kan vokabularene dra nytte av eksterne ressurser som eksisterer som RDF/ linked open data?
- Kan Realfagstermer dra nytte av TEKORDS hierarkiske struktur?

1.3 Prosjektgruppen

Prosjektgruppen er sammensatt av representanter fra både Realfagsbiblioteket og NTNU UB. Samlet har gruppen kompetanse bl.a. innenfor emneord, klassifikasjon, biologi, kjemi, informatikk, programmering og semantisk web:

Ellen Samdahl Flatby (universitetsbibliotekar) (NTNU UB)

Rurik Greenall (freelance utvikler)

Knut Hegna (førstebibliotekar, Realfagsbiblioteket)

Mari Lundevall (spesialbibliotekar, Realfagsbiblioteket)

Kyrre Traavik Låberg (overingeniør, Realfagsbiblioteket)

Heidi Sjursen Konestabo (førstebibliotekar, Realfagsbiblioteket)

Viola Kuldvere (prosjektleder/spesialbibliotekar, Realfagsbiblioteket)

Gruppen har hatt noen felles møter og ellers kommunisert per e-post og Skype. To representanter har deltatt på 2 samlinger med «Linked data – et nettverk» (NTNU UB) høsten 2012.

5 representanter har vært på konferansen Beyond Libraries: Subject Metadata in the Digital Environment and Semantic Web - IFLA Satellite Post-Conference i Tallinn (17.-18. august 2012).

Vi har presentert prosjektet på Seminar om mapping til Dewey ved HiOA (11. juni 2012) og på Kunnskapsorganisasjonsdagene (KORG) 2013.

2. Datagrunnlag

TEKORD

TEKORD er en kontrollert, hierarkisk emneordsliste som er utviklet ved NTNU UB og brukes for de tekniske og naturvitenskapelige fagene.

Emneordslisten het opprinnelig NTNUB's emneordsliste, og det var hovedbibliotekar Ansteinson som trakk opp retningslinjene for emneordslisten en gang i 1920-årene [10]. Det ble lagt vekt på at emneordslisten skulle inneholde norske ord. NTNUB's emneordsliste eksisterte i bibliotekets kortkatalog fram til innføringen av BIBSYS i 1980. Da ble emneordslisten overført til BIBSYS emnemodul og slik eksisterer den fremdeles.

I dag inneholder TEKORD ca. 15 000 termer. Listen inneholder overordnede og underordnede termer (OT og UT), se-henvisninger og se også-henvisninger. Hvert emneord er koblet til en UDK-klassekode.

Det er en liten tilvekst av termer per år. Det blir tilført nye emneord når nye fagområder kommer til ved NTNU. Fagansvarlig innen hvert fagområde tilfører emneordene og finner tilhørende UDK-klassifikasjon fra UDC online [11]. Fagansvarlige har kontakt med de enkelte fagmiljø ved NTNU, slik at emneordene blir i overensstemmelse med fagterminologien. Vi vet at enkelte norske høyskolebibliotek som har tekniske fag, bruker TEKORD. Det finnes per i dag ingen oversikt over hvilke bibliotek dette er.

Realfagstermer

Realfagstermer er et kontrollert vokabular som er utarbeidet av Realfagsbiblioteket. Utgangspunktet er frie nøkkelord som er brukt på bøker i bibliotekets samlinger. Vokabularet er gruppert i indekstermer, formtermer, stedtermer og tidtermer. Alle de fire typene kan ha synonymer, oversettelser og/eller akronymer. I tillegg kan det være registrert klassifikasjon (DDC eller MSC), noter (til internt bruk) og definisjoner (tiltenkt sluttbruker).

I materialet som er brukt i prosjektet har vi 15 200 termer.

Av disse har:

1 592 engelsk oversettelse,

141 latin,

292 DDC-klassifikasjon.

På indekstermene er det dessuten fordelt:

2 076 synonymer,

460 se også-henvisninger,

632 strenger.

Parallelt med prosjektet har vi gjort kontinuerlige rettelser og oppdateringer i Realfagstermer. Vokabularet i dag er derfor ikke identisk med det som er brukt i sammenlikningene.

Realfagsbiblioteket har også utarbeidet et websøk som bruker Realfagstermer for oppslag i samlingene gjennom BIBSYS. Websøket gir treff på foretrukne termer, også ved søk på synonymer. I tillegg gis det lenker til oppslag i norsk og engelsk Wikipedia, Store Norske Leksikon, Periodesystemet, MSC og DDC. Websøket brukes også av andre vokabularer, som TEKORD, Senter for menneskerettigheter og HUMORD [12].

3. Metoder

Prosjektet ønsket en komparativ analyse utført på den terminologiske overlappen mellom TEKORD og Realfagstermer. Første skritt var å finne denne overlappen. Både for å finne overlappen og for å sammenlikne videre der vi har en overlapp i terminologi, har vi i hovedsak ønsket å teste ut en kvantitativ og automatisk metode der vi bl.a. er interessert i i hvilken grad den automatiske metoden gir gode resultater eller må suppleres med manuelle/intellektuelle studier av data.

Grunnlaget for å kunne gå løs på en sammenlikning av TEKORD og Realfagstermer var å bruke RDF som plattform. TEKORD eksisterte allerede i RDF-format, strukturert i Simple Knowledge Organization System (SKOS) [13] og var altså åpent tilgjengelig til å brukes av interesserte. Realfagstermer ble i første fase av prosjektet konvertert til SKOS/RDF slik at begge vokabularene forelå som åpne data i et sammenlignbart format.

RDF som teknologi gir noen gunstige fordeler. RDF gir mulighet for å lage spørringer med et felles spørrespråk (SPARQL [14]) mot begge datasett. Det gir også mulighet for å samkjøre dataene med andres data. I tillegg blir dataene åpne og tilgjengelige via en brukbar protokoll (http). SKOS gir mulighet for å strukturere emneord i relasjoner og hierarki innenfor RDF-syntaksen. Begrepene (SKOS Concept) får en unik ID.

Vi valgte å prøve SILK Link Discovery Framework [15], som er et system som kan brukes for å sammenlikne to vokabularer med lik syntaks (som vokabularene får med RDF/SKOS) ved hjelp av en gitt sammenlikningsmetode. Etter prøving av ulike sammenlikningsalgoritmer ble distanse-algoritmen Jaro [16] valgt. Jaro regner ut graden av likhet mellom to tekststrenger. Likhetsgraden mellom tekststrengene kan justeres.

I sammenlikningen av de to datasettene i RDF valgte vi å finne fram til de emneordene som har en tekstlikhetsgrad på mellom 90 – 100 %, for å fange opp ord som kan ha lik betydning, men noe ulik skrivemåte (for eksempel ulik endelse).

Siden den maskinelle sammenlikningen finner tekstlikhet mellom ord, betyr det at vi i utgangspunktet sammenliknet *termer* og ikke *begreper*. Det ble kjørt en sammenlikning av alle termene i hvert vokabular (både foretrukne termer og synonymer til disse) mot hverandre. Det betyr at en foretrukken term i det ene vokabularet kan ha tekstlikhet med et synonym i det andre vokabularet og disse blir da listet som like ordpar.

Listen vil da også vise en kobling mellom de foretrukne termene i begge vokabularer til tross for at de ikke nødvendigvis er tekstlike, for eksempel:

TEKORD	**Realfagstermer**
Biologisk mangfold (T)	Biologisk mangfold (BF)
Biologisk mangfold (T)	Biodiversitet (T)

(T = foretrukken term. BF = brukt for/synonym)

Den maskinelle sammenlikningen ga oss flere lister med emneordspar. Det var nødvendig med en manuell gjennomgang av deler av materialet for å luke ut termer som den maskinelle sammenlikningen feilaktig hadde listet som like (med lik betydning), for eksempel homonymer. Metoden har på den annen side den fordel at synonymer kan fanges opp, og ved å justere likhetsgraden mellom termene gir den oss mulighet til manuelt å finne ordene som har lik betydning, men noe ulik skrivemåte.

Vi var videre interessert i hvor mange, og hvilke *begreper* (SKOS Concepts) som fantes i begge vokabularene. Disse ble identifisert ved hjelp av SKOS-identifikatorer og kjørt ut i en ny liste kalt Concepts. Disse begrepene med tilhørende synonymer, språk, relasjoner m.m. ser vi nå på som Realfagstermers og TEKORDs felles univers, den overlappen vi ønsket å finne. Dette ble vårt videre forskningsmateriale med tanke på mulig deling og berikelse.

Det ble gjort videre spørringer med SPARQL mot RDF-filene, denne gangen begrenset til overlappen, for å prøve å finne svar på våre problemstillinger rundt deling og berikelse.

For å undersøke om TEKORD og Realfagstermer har ulike synonymer og assosiative relasjoner/se også-henvisninger som kan deles, ble følgende spørsmål utarbeidet som utgangspunkt for SPARQL-spørringene mot RDF-filene:

- Hvor mange av Realfagstermers synonymer (SKOS AltLabel) finnes som foretrukne termer (SKOS PrefLabel) i TEKORD?
- Hvor mange av Realfagstermers foretrukne termer finnes som synonymer i TEKORD?
- Hvor mange av TEKORDs synonymer finnes som synonymer i Realfagstermer?
- Hvor mange av Realfagstermers synonymer finnes som synonymer i TEKORD?
- Hvor mange av Realfagstermers se også-henvisninger (SKOS Related terms) finnes som synonymer eller foretrukne termer i TEKORD - uten at den samme relasjonen allerede er uttrykt i TEKORD?
- Hvor mange av TEKORDs se også-henvisninger finnes som synonymer eller foretrukne termer i Realfagstermer - uten at den samme relasjonen allerede er uttrykt i Realfagstermer?

Vi ville også finne ut om TEKORD kunne få nytte av Realfagstermers engelske emneord:

- Hvor mange felles foretrukne termer i Realfagstermer og TEKORD har engelsk oversettelse i Realfagstermer?

I tillegg var det et mål å prøve ut om vokabularene kunne dra nytte av eksterne ressurser som eksisterer som RDF. Vi valgte i første omgang DBpedia [17] som ressurs, med formål å se etter engelske termer som samsvarer med våre norske i både TEKORD og Realfagstermer:

- Hvor mange foretrukne termer i TEKORD og Realfagstermer får kobling til engelsk DBpedia-term?

Til slutt ville vi vite om Realfagstermer kunne dra nytte av TEKORDS hierarkiske struktur:

- Hvor mange foretrukne termer i TEKORD har underordnet term/UT (SKOS Narrower term) eller overordnet term/OT (SKOS Broader term)?
- Hvor mange av disse underordnede og overordnede termene finnes også i Realfagstermer?

For disse problemstillingene fikk vi ut lister som beskrives nærmere i kapittel 4.

4. Resultater og diskusjon

4.1 Terminologisk overlapp mellom TEKORD og Realfagstermer

Den automatiske sammenlikningen på termnivå for å finne emnesystemenes terminologiske overlapp, resulterte i 4 lister med ulik grad av tekstlikhet mellom emneordene:

ExactMatches > 95 % likhet (=3 473 termer)

NearMatches 90-95 % likhet (=920 termer)

NomatchesRT (termer som finnes bare i Realfagstermer)

NomatchesTO (termer som finnes bare i TEKORD)

Som før nevnt, var ikke alle termer i lista ExactMatches helt like ettersom den automatiske sammenlikningen hadde et nøyaktighetskriterium på 95 % eller mer. Termene som ikke hadde 100 % ordlikhet ble skilt ut i en egen liste og gjennomgått manuelt for å luke vekk feilkoblinger. NearMatches (de nesten like) gikk vi også igjennom manuelt. Vi aksepterte termer med lik betydning med J og forkastet ulike (termer med ulik betydning) med N. De *virkelige* ExactMatches, altså de som hadde 100 % tegnlikhet, ble automatisk akseptert som like.

Tabell 1: Eksempler fra ExactMatches (manuell gjennomgang)

Realfagstermer	**TEKORD**	
Alkener	Olefiner	J
Almanakker	Kalendere	J
Alternativ medisin	Naturmedisin	J
Aluminiumframstilling	Aluminiumfremstilling	J
Andre verdenskrig	Verdenskrigen 1939-1945	J
Application service provider	ASP	N
Arbeider	Arbeidere	J
Arbeidsmedisin	Yrkesmedisin	J
Årevinger	Veps	J
Askorbinsyre	Vitamin C	J
Bakterier	Bakerier	N

Tabell 2: Eksempler fra NearMatches (manuell gjennomgang)

Realfagstermer	**TEKORD**	
Bærekonstruksjoner	Rørkonstruksjoner	N
Bæreteknikk	Treteknikk	N
Banach algebra	Banach-algebraer	J
Barkbiller	Briller	N
Barnesykdommer	Øresykdommer	N
Barriere	Karriere	N
Basalter	Basalt	J
Bedriftsadministrasjon	Bedriftsadministrasjon - Etikk	N

Etter den manuelle gjennomgangen av ordparene i listene endte vi opp med 3 397 ordpar som vi aksepterte som like (med lik betydning) og 996 ordpar som vi definerte som ulike (med ulik betydning).

Denne første maskinelle sammenlikningen som dreide seg om ordlikhet ble altså gjort på alle termene i hvert vokabular (både foretrukne termer og synonymer til disse). Denne metoden har, som vi har sett, sin begrensning i og med at den ikke fanger opp homonymer. Synonymer fanges opp bare dersom en term som er lik nok til å fanges opp er brukt i begge vokabularer.[1] I tillegg krever metoden intellektuell kontroll for betydningslikhet mellom ordpar som ikke har helt lik tekststreng. I og med at vi har akseptert den listen som hadde 100 % tekstlikhet mellom ordparene uten å ta en manuell sjekk, har vi sannsynligvis fått med oss noe støy i form av homonymer som ikke er fanget opp.

For *begreper* (SKOS Concept) identifiserte vi en kobling for 3 243 begreper. Det vil altså si en kobling som sier at et begrep i TEKORD er det samme som et begrep i Realfagstermer. Denne terminologiske overlappen mellom TEKORD og Realfagstermer utgjør ca. 20 % av vokabularene.

Tabell 3: Eksempler fra SKOS Concept (overlapp på begrepsnivå)

Realfagstermer	**TEKORD**
Fjellplanter	Fjellflora
Samfunnsutvikling	Samfunnsutvikling
Svangerskap	Svangerskap
Kromatografi	Kromatografi
Byhistorie	Byhistorie
Tekstbehandling	Tekstbehandling
Biodiversitet	Biologisk mangfold
Selfangst	Selfangst

Nå vil noen kanskje mene at en terminologisk overlapp på 20 % er en liten grad av likhet med tanke på at vi har omtrent 15 000 termer i hvert vokabular. Samtidig kan vi tenke oss at en for stor grad av likhet/overlapp

[1] Som eksemplet i Tabell 1: Alkener (Realfagstermer) og Olefiner (TEKORD) er fanget opp som treff, fordi TEKORD har angitt at Olefiner er synonym til den foretrukne termen Alkener.

heller ikke ville vært et godt utgangspunkt for deling og utveksling. Overlappen finner vi der vi har likhet i våre fagsamlinger. Noe av målet vil være å samordne og tilføre hverandre noe innen for dette felles universet.

4.2 Synonymer og assosiative relasjoner

I spørsmålet om muligheten for å dele synonymer (SKOS AltLabels) ga spørringene mot RDF-filene innenfor vokabularenes terminologiske overlapp oss nye filer/lister med vokabularenes SKOS-ID og termer.

Det var 138 foretrukne termer i Realfagstermer som fantes som synonymer i TEKORD. Potensielt kan TEKORD tilføre Realfagstermer synonymer (TEKORDs foretrukne term og eventuelle tilleggssynonymer knyttet til denne kan overføres).

Vi fikk 158 termer i TEKORD som forekom som synonymer i Realfagstermer. Potensielt kan Realfagstermer tilføre TEKORD synonymer (Realfagstermers foretrukne term og eventuelle tilleggssynonymer knyttet til denne kan overføres).

I 39 tilfeller finnes et synonym (SKOS AltLabel) i TEKORD også samtidig som synonym i Realfagstermer.

Eksempler - synonymer

Eksemplene viser TEKORDs og Realfagstermers bruk av foretrukne termer (T) og synonymer (Brukt for/BF). Eksemplene framkom etter manuell kontroll (koblingsordet er understreket):

TEKORD	**Realfagstermer**
T: Kalendere	T: Almanakker
BF: Almanakker	

Realfagstermer kan her få et synonym fra TEKORD (Kalendere).

TEKORD	**Realfagstermer**
Term: Dammer	Term: Demninger
BF: Demninger	
BF: Damanlegg	
BF: Dambygging	

Realfagstermer kan her potensielt få 3 synonymer fra TEKORD, men brukt for-termene som TEKORD har her er ikke virkelige synonymer. Disse burde kanskje heller vært se også-henvisninger. Dette er et eksempel på at man ved emneordsindeksering har en pragmatisk, samlingsspesifikk tilnærming til bruk av se-henvisninger.

TEKORD	**Realfagstermer**
T: Brannvegger	T: Brannmur
	BF: Brannvegger

Potensielt kan Realfagstermer tilføre TEKORD et synonym (Brannmur), men det viser seg at «Brannvegger» i TEKORD og «Brannvegger» i Realfagstermer ikke betegner det samme. Termen har byggteknisk betydning i TEKORD og datateknisk i Realfagstermer. Om vokabularene gjennomgående hadde benyttet kvalifikatorer som angir fagområde, kunne denne formen for feilkoblinger kanskje vært unngått.

TEKORD	**Realfagstermer**
T: Elektrisitetsforsyning	T: Elektrisitetsproduksjon
	BF: Elektrisitetsforsyning
	BF: Elektrisitetsverk

Her kunne man tenke seg at Realfagstermer kan tilføre TEKORD to synonymer. Men ved en nærmere sjekk viser det seg at TEKORD har Elektrisitetsverk som egen foretrukken term med Elektrisitetsproduksjon og Kraftstasjoner som synonymer til denne.

TEKORD	**Realfagstermer**
T: Arktiske strøk	T: Polarområder
BF: Polare strøk	BF: Polare strøk
BF: Polarstrøk	BF: Polområder
BF: Kalde strøk	

Eksempelet der det er funnet en kobling mellom emnesystemenes synonymer, viser at deling og samordning er mulig, men alle mulige koblinger må undersøkes manuelt.

For *assosiative relasjoner* (se også-henvisninger) er det ikke produsert noen liste. Det viste seg at TEKORD ikke har representert slike relasjoner (SKOS Related terms) som RDF. Vi har derfor ikke gått videre med disse.

4.3 Engelske termer fra Realfagstermer til TEKORD

Filen viser 439 forekomster der TEKORD og Realfagstermer har samme foretrukne term hvor Realfagstermer også har en engelsk oversettelse av termen. Disse kan berike TEKORD og gi en engelskspråklig inngang til deler av vokabularet. Kilden for de engelske termene i Realfagstermer er i hovedsak amerikanske emneordssystemer for henholdsvis matematikk og informatikk, så kvaliteten her er god.

Tabell 4: Eksempler på engelsk fra Realfagstermer til TEKORD

Foretrukken term i TEKORD og Realfagstermer	**Engelsk oversettelse i Realfagstermer**
IDL	Interface definition language
Virveldyr	Vertebrates
Elliptiske kurver	Elliptic curves
Fysiologi	Physiology
Boolsk algebra	Boolean algebra
Datalingvistikk	Computational linguistics
Dataanalyse	Data analysis
Industri	Industry
Kommunikasjonsprotokoller	Communication protocols
Målteori	Measure; measure theory
Planettåker	Planetary nebulae
Termodynamikk	Thermodynamics

4.4 Engelske termer fra DBpedia til TEKORD og Realfagstermer

I spørsmålet om hvor mange foretrukne termer i TEKORD og Realfagstermer som får en DBpedia-kobling (DBpedia-lenke (URL)) med engelsk oversettelse, fikk vi 2 111 unike DBpedia-lenker i listen som ble generert. I 1 045 tilfeller har både TEKORD og Realfagstermer fått kobling mot en (og samme) DBpedia-lenke. I 121 tilfeller har TEKORD alene fått en kobling mot en DBpedia-lenke, mens Realfagstermer alene har 945 koblinger mot en DBpedia-lenke.

1 241 DBpedia-lenker er fordelt på 458 unike TEKORD-ID'er (SKOS-ID). 2 090 DBpedia-lenker er fordelt på 662 unike Realfagstermer-ID'er (SKOS-ID).

Mange koblinger mellom våre norske termer og DBpedia-lenker synes å være tilfredsstillende.

Eksempel: Holografi

Realfagstermers "Holografi" er sendt til DBpedia:

<skos:Concept rdf:about="#REAL07560"><rdf:type
rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/><skos:PrefLabel
xml:lang="no">Holografi</skos:PrefLabel></skos:Concept>
<http://www.w3.org/2002/07/owl#sameAs>http://DBpedia.org/page/Holography

Søket gir treff på den engelske oversettelsen "Holography".

En unik Realfagsterm-ID eller TEKORD-ID har, som vi ser av tallene ovenfor, i mange tilfeller fått kobling til flere DBpedia-lenker. Disse lenkene representerer ulike forslag til kobling til engelsk term der TEKORD/Realfagstermer har et emneord som er flertydig i DBpedia.

Eksempel: Astronomi

REAL16547,<http://DBpedia.org/resource/Category%3AAstronomy> .
REAL16547,<http://DBpedia.org/class/yago/Astronomy(DragonlandAlbum)> .
REAL16547,<http://DBpedia.org/class/yago/Astronomy(song)> .
REAL16547,<http://DBpedia.org/class/yago/Astronomy(BleachAlbum)> .
REAL16547,<http://DBpedia.org/class/yago/AstronomyMagazine> .

Eksempel: Energi

REAL11549,<http://DBpedia.org/resource/Category%3AEnergy> .
REAL11549,<http://DBpedia.org/class/yago/Energy(PointerSistersAlbum)>
REAL11549,<http://DBpedia.org/class/yago/Energy(esotericism)> .
REAL11549,<http://DBpedia.org/class/yago/Energy(album)> .
REAL11549,<http://DBpedia.org/class/yago/Energy,Illinois> ."
REAL11549,<http://DBpedia.org/class/yago/Energy(jazzAlbum)>
Osv.

Det finnes også eksempler på feilkoblinger:

Eksempel: Maur

REAL08928,<http://DBpedia.org/resource/Maur,Switzerland>

Termen "Maur" er sendt til DBpedia, og returnerer ett treff: stedsnavnet
Maur i Sveits og ikke "Ants" som ønsket.

Disse koblingene skaper en god del støy og må altså gjennomgås manuelt. Dette tyder på at metoden for kobling mellom termene i vokabularene og DBpedia har forbedringspotensiale. Det kan også være en idé å sammenlikne emneordene i Realfagstermer som har engelsk oversettelse fra før, med forslagene fra

DBpedia for å se på i hvilken grad det er samsvar.

Å koble de to vokabularenes RDF-filer mot DBpedia viste seg også å være en ganske omfattende jobb på grunn av den store datamengden som skulle leses og returnere svar.

4.5 Hierarki

Vi ønsket å finne ut om det var grunnlag for at Realfagstermer kunne bruke TEKORDs hierarki til å bygge hierarkiske relasjoner mellom egne termer. Vi undersøkte hvor mange foretrukne termer i TEKORD som hadde underordnet term/UT eller overordnet term/OT. I tillegg ville vi se hvor mange av disse overordnede og underordnede termene i TEKORD som også finnes i Realfagstermer.

TEKORD har 6 160 OT og 6 164 UT. Innenfor de to vokabularenes overlappende (foretrukne) termer har 1 560 enten OT eller UT.

De fordeler seg slik:

115 termer i Realfagstermer kan potensielt tilføres både OT- og UT-relasjon i sitt eget vokabular.

342 termer i Realfagstermer kan potensielt tilføres UT-relasjon i sitt eget vokabular.

696 termer i Realfagstermer kan potensielt tilføres OT-relasjon i sitt eget vokabular.

637 Overlappende termer har ikke UT/OT i Realfagstermer.

Eksempler på at Realfagstermer kan bygge hierarkiske relasjoner basert på strukturen i TEKORD:

Både OT og UT-relasjon:

Akustikk kan få OT Fysikk, og UT Undervannsakustikk, Bygningsakustikk, Romakustikk og Elektroakustikk

UT-relasjon:

Grafikk kan få UT Litografi

Variable stjerner kan få UT Pulsarer

OT-relasjon:

Reinsdyr kan få OT Hjortedyr

Svarte hull kan få OT Kosmologi

Prosjektgruppen synes dette ser lovende ut, men også når det gjelder hierarki, må termer og relasjonene mellom dem kontrolleres manuelt.

Som vist ovenfor, gir listene oss først og fremst et kvantitativt resultat som svar på våre problemstillinger. Listene gir ingen visualisering av for eksempel TEKORDs hierarkiske struktur eller hvilke synonymer som kan overføres mellom vokabularene. For å kunne si noe kvalitativt om resultatet og berikelsesmuligheter, var det nødvendig med en manuell sjekk av et utvalg termer og termenes sammenheng/omgivelser (som vist i eksemplene). Analyse av resultater har derfor vært tidkrevende.

Å utnytte ulike språkvarianter ser ut til å kunne gjøres med en høyere grad av automatikk, mens det vil være krevende å gå videre med å dele synonymer og hierarki. Det ville være en fordel om vi i beskrivelsen av vokabularene skilte mellom se-henvisninger som virkelig er synonymer, og se-henvisninger som er laget utfra en vurdering om at det er hensiktsmessig i det enkelte vokabularet. I tillegg vil bruk av kvalifikatorer på homonymer være svært nyttig.

5. Oppsummering og konklusjon

De to emnesystemene TEKORD og Realfagstermer er blitt sammenliknet med RDF/SKOS som plattform. Vi testet en maskinell/automatisk metode (SILK og Jaro) som gikk ut på å sammenlikne termers tekstlikhet for å finne en terminologisk overlapp mellom vokabularene med tanke på mulig berikelse, samordning og utveksling av emneord. Vokabularenes overlappende *begreper* ble så koblet ved hjelp av SKOS-identifikatorer.

Berikelsesmuligheter for vokabularene

Vi fant en terminologisk overlapp på ca. 20 % av de to sammenliknede emnesystemene. Overlappen vil naturlig nok være der det er sammenfall mellom fagsamlingene og dermed også mellom terminologien. Dette vil gjelde først og fremst for det matematisk-naturvitenskapelige området. TEKORD har sin overvekt av emner innenfor teknikk, mens Realfagstermer har overvekten på de naturvitenskapelige emnene. Prosjektgruppen mener dette er et godt utgangspunkt for videre studier.

TEKORD og Realfagstermer kan i rundt 300 tilfeller potensielt utveksle synonymer. At vokabularene har en noe ulik praksis for emneordssetting og i flere tilfeller har gjort ulike valg når det gjelder foretrukne termer og synonymer til disse, gir muligheter for å berike hverandres vokabular ved deling av synonymer. Samtidig viser det seg at kvalitetskontroll er nødvendig før en deling på grunn av pragmatisk bruk av henvisninger.[2] Emneordspraksisen er med andre ord ikke konsekvent. Assosiative relasjoner har vi foreløpig valgt å se bort i fra fordi disse ikke er representert som RDF i TEKORD.

Realfagstermer kan berike TEKORD med 439 engelske termer av god kvalitet. Dette kan gjøres uten noen stor manuell sjekk. Neste skritt blir å finne en metode for overføring av termene.

Når det gjelder spørsmålet om vokabularene kan dra nytte av eksterne ressurser som eksisterer som linked open data, er den foreløpige konklusjonen etter et forsøk på å koble vokabularenes felles norske termer mot engelske termer i DBpedia, at det ikke er noen automatisk og rask måte å få til dette på. Mye støy på grunn av tvetydige termer som får mange irrelevante koblingsforslag og en del feilkoblinger medfører mye manuelt

[2] Synonymer er egentlig se også-henvisninger, homonymer er ikke alltid fanget opp, et synonym til en foretrukken term i det ene vokabularet kan være en egen foretrukken term i det andre vokabularet m.m.

arbeid med kvalitetssjekk. Eventuelt må metoden forbedres. Vi ser allikevel også mange nyttige koblinger som etter en videre sjekk kan tas inn i vokabularene og slik gi en engelskspråklig inngang til mange emner.

Vi har funnet at potensielt kan i overkant av 1 000 av Realfagstermene få tilført overordnede og/eller underordnede termer ved å bruke TEKORDs hierarkiske struktur som utgangspunkt til å bygge hierarkiske relasjoner mellom egne termer. Her må det også en gjennomgående kvalitetskontroll til før en hierarkisk struktur kan ta form, men det er grunn til å være optimistisk på bakgrunn av stikkprøver. Det må videre utredes om, og i tilfelle hvordan, strukturen skal tas inn i vokabularet.

RDF som plattform for sammenlikning og eventuell sammenføyning av vokabularer

En forutsetning for at TEKORD og Realfagstermer kunne sammenliknes automatisk var at begge vokabularene var tilgjengelige i et sammenliknbart format hvor man kan bruke et felles spørrespråk.

Prosjektgruppen mener at RDF har fungert godt som inngangsport til en automatisk sammenlikning mellom vokabularene. I RDF kombinert med med SILK- og Jaro- metodene har vi funnet en maskinell/automatisk sammenlikningsmetode som kan føre et stykke på veg i sammenlikningen av emnesystemer. Metoden gir først og fremst kvantitative resultater som så må analyseres videre med intellektuelle metoder. Vi har altså ikke funnet en rask snarvei til å koble to eller flere emnesystemer sammen, men en metode som kan fungere som utgangspunkt for sammenlikning.

Vi ser per i dag ikke at det er hensiktsmessig å slå TEKORD og Realfagstermer sammen til ett felles vokabular, da den største andelen av termene i de to vokabularene tilhører ulike fagområder, og det vil kreve en innsats som ikke vil stå i forhold til resultatet. Vi mener det vil være mer nyttig å heve kvaliteten på den terminologiske overlappen som vi har funnet ved videre samarbeid om å berike vokabularene med deling og eventuell samordning av termer, der vi har funnet at dette lar seg gjøre. Det kan også bli aktuelt å utvikle et søkegrensesnitt for overlappen mellom TEKORD og Realfagstermer for bibliotekarer og sluttbrukere slik at den kan utnyttes i henholdsvis emneordsindeksering og søking etter litteratur.

Det kunne være mulig i framtiden å finne overlapp i terminologi også til andre vokabularer (for eksempel Agrovoc [18]) og slik bygge videre på det vi har gjort nå slik at flere vokabularer gjensidig kan berikes og kvalitetsheves. Med mer erfaring kan også metodene for sammenlikning og deling utvikles. Det går også an å tenke seg at det legges til rette for et felles grensesnitt der ulike vokabularer søkes i som ett, og der treff vises med beholdningsinformasjon. På denne måten er det mulig å samle mange vokabularer. En slik løsning mener vi det er grunnlag for at Nasjonalbiblioteket ser nærmere på. Forslaget bør også tas med i betraktning når det gjelder nytt biblioteksystem for BIBSYS-bibliotekene.

Referanser

1. RDF Working Group. Resource description framework (RDF): W3C; [2013-05-02]. URL: http://www.w3.org/RDF/.

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

3. W3C. Linked data: W3C; [2013-05-02]. URL: http://www.w3.org/standards/semanticweb/data.

4. Semantic Computing Research Group (SeCo). National Semantic Web Ontology Project in Finland (FinnONTO), 2003-2012 [2013-05-03]. URL: http://www.seco.tkk.fi/projects/finnonto/.

5. Lov av 1. april 2005 nr. 15 om universiteter og høyskoler (Universitets- og høyskoleloven) : med endringer, sist ved lov av 22. juni 2012 nr. 55. Oslo: Kunnskapsdepartementet; 2005.

6. American Mathematical Society. MSC2010 database 2010 [2013-05-02]. URL: http://www.ams.org/mathscinet/msc/msc2010.html.

7. OCLC. WebDewey [2013-05-06]. URL: http://dewey.org/webdewey/.

8. UiO: Universitetsbiblioteket. Emneordsøk mot BIBSYS : [realfag] [2013-05-02]. URL: http://app.uio.no/ub/emnesok/?id=UREAL.

9. Universitetet i Oslo. Kjemisk institutt. Periodesystemet.no [2013-05-06]. URL: http://www.mn.uio.no/kjemi/tjenester/kunnskap/periodesystemet/.

10. Lomheim I. Emneordklassifisering og terminologiutvikling ved NTUB. NTUB-75 år : Norges tekniske universitetsbibliotek 1912-1987. [Trondheim]: Tapir; 1987. S. 65-9.

11. British Standards Institution. UDC Online [2013-05-03]. URL: http://www.udconline.net/.

12. UiO: Universitetsbiblioteket. Emneordsøk mot BIBSYS [2013-05-02]. URL: http://app.uio.no/ub/emnesok/.

13. W3C Semantic Web Deployment Working Group (SWDWG). SKOS Simple Knowledge Organization System: W3C; 2009 [2013-05-02]. URL: http://www.w3.org/2004/02/skos/.

14. W3C. SPARQL query language for RDF : W3C recommendation [2013-05-02]. URL: http://www.w3.org/TR/rdf-sparql-query/.

15. Isele R, Jentzsch A, Bizer C, Volz J. Silk : a link discovery framework for the web of data Berlin: Freie Universität Berlin; [2013-05-02]. URL: http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/.

16. Jaro-Winkler distance [2013-05-07]. URL: http://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance.

17. DBpedia [2013-05-02]. URL: http://dbpedia.org/.

18. AGROVOC [2013-05-08]. URL: http://aims.fao.org/standards/agrovoc/.