Nasjonalbiblioteket

UiO : Universitetsbiblioteket

En norsk generell tesaurus?

Sluttrapport med anbefalinger fra Tesaurus forprosjekt (5.3.2014-1.3.2015)

Oddrun Pauline Ohren[1], Dan Michael Heggø[2], Berit Sonja Hougaard[2], Lars Johnsen[1], Unni Knutsen[2], Lembi Viola Kuldvere[2], Vibeke Stockinger Lundetræ[2], Mari Lundevall[2], Kirsten Rydland[1], Ingebjørg Rype[1] og Torstein Tjelta[1]

[1]Nasjonalbiblioteket
[2]Universitetsbiblioteket i Oslo

Oslo 2. mars 2015

Innhold

1 Innledning

Denne rapporten sammenfatter resultatene fra Tesaurus forprosjekt. Hensikten med forprosjektet har vært å utrede hva det innebærer å etablere en generell tesaurus (heretter *Norsk generell tesaurus* eller *NGT*) basert på den eksisterende tesaurusen Humord. Arbeidet i forprosjektet er beskrevet i prosjektbeskrivelsen (se Appendiks 1), som oppsummerer de konkrete resultatmålene slik:

«Forprosjektet skal resultere i:

1. *Plan for utvikling av Norske emneord*[1] *versjon 1.0 på bokmål, inkludert aktiviteter, tidsplan med milepæler, deltakere/organisering og ressursestimat.*
2. *Forslag til driftsmodell for Norske emneord*
3. *Norske Emneord versjon 0.1 på bokmål*
4. *«Veikart» for Norske emneord: Forslag til og plan for videreutvikling: …»*

I det følgende er de fire delresultatene beskrevet etter tur, og skal til sammen danne beslutningsgrunnlag i spørsmålet om *hvorvidt* det skal etableres en generell tesaurus og eventuelt *hvordan* utvikling av en slik bør gripes an.

Prosjektgruppen har bestått av følgende medlemmer:

- Oddrun Pauline Ohren, Nasjonalbiblioteket (prosjektleder)
- Dan Michael Heggø, Universitetsbiblioteket i Oslo, Realfagsbiblioteket
- Berit Sonja Hougaard, Universitetsbiblioteket i Oslo, HumSam-biblioteket
- Lars Johnsen, Nasjonalbiblioteket
- Unni Knutsen, Universitetsbiblioteket i Oslo, HumSam-biblioteket
- Lembi Viola Kuldvere, Universitetsbiblioteket i Oslo, Realfagsbiblioteket
- Vibeke Stockinger Lundetræ, Universitetsbiblioteket i Oslo, HumSam-biblioteket
- Mari Lundevall, Universitetsbiblioteket i Oslo, Realfagsbiblioteket
- Kirsten Rydland, Nasjonalbiblioteket
- Ingebjørg Rype, Nasjonalbiblioteket
- Torstein Tjelta, Nasjonalbiblioteket

1.1 Rapportens struktur

Rapporten er strukturert som følger:

Kapittel 2 oppsummerer prosjektets anbefalinger når det gjelder felles løsning for emneordssystem.

[1] Prosjektbeskrivelsen bruker «Norske emneord» som navn på den generelle tesaurusen. I løpet av forprosjektet ble dette forlatt til fordel for «Norsk generell tesaurus» (forkortet «NGT»), som følgelig brukes i denne rapporten.

Kapittel 3 beskriver kort tidligere relevant arbeid i Nasjonalbiblioteket og Universitetsbiblioteket i Oslo, og institusjonenes motivasjon for å delta i Tesaurus forprosjekt.

Kapittel 4 gjør rede for prosjektets kommunikasjon med bibliotekfeltet og konklusjonene som er trukket på basis av dette.

Kapittel 5 beskriver hva NGT skal inneholde, generelt og for første versjon (NGT 1.0) spesielt.

Kapittel 6-10 utgjør prosjektplan for utvikling av NGT 1.0, inkludert aktivitets- og tidsplan, prosjektorganisasjon, infrastruktur og finansieringsforslag.

Kapittel 11 beskriver forslag til modell for drift av NGT, fra versjon 1.0 av.

Kapittel 12 skisserer retning for NGTs utvikling etter første versjon (NGT 1.0).

Rapporten inneholder også appendikser, som hovedsakelig gir utfyllende informasjon til kapitler i selve rapporten.

I resten av denne rapporten står *NB* for *Nasjonalbiblioteket* og *UBO* for *Universitetsbiblioteket i Oslo*.

2 Anbefalinger

Her oppsummeres de viktigste anbefalingene fra *Tesaurus forprosjekt*. Forslagene er utdypet i de andre delene av rapporten.

2.1 Visjon

Visjonen er å komme fram til et nasjonalt emneordssystem

- som er generelt i den forstand at den dekker et bredt fagfelt.
- som er på tesaurusform som lenkede data og knyttet opp mot det øvrige nettverk av emneautoriteter.
- som skal inneholde hovedsakelig innholdsbeskrivende begreper
- hvor begrepene er egnet til bruk for indeksering i fag- og forskningsbibliotek.
- hvor alle begreper har termer på bokmål, nynorsk og engelsk, og så mange som mulig har termer på samiske språk og kvensk.
- som er åpent tilgjengelig for fri bruk av alle.
- som forvaltes distribuert av fagmiljøene, koordinert av en sentral redaksjon.

Foreløpig går emneordssystemet under navnet *Norsk generell tesaurus (NGT)*, se kapittel 5 for nærmere beskrivelse.

NGT vil bli en verdifull ressurs for kunnskapsorganisering så vel som terminologiarbeid. Felles forvaltning vil effektivisere ressursbruken i bibliotekene både når det gjelder emneordsutvikling og indeksering. I tillegg øker muligheten for å utvikle gode digitale sluttbrukertjenester. Som intellektuelt kuratert emnestruktur

vil NGT også danne et godt utgangspunkt for språkteknologiske anvendelser, eksempelvis automatisk innholdsanalyse av fulltekst.

2.2 Veien fram

Vi foreslår en stegvis tilnærming, og anbefaler i første omgang å initiere et samarbeidsprosjekt mellom NB og UBO som har som mål utvikle første versjon av NGT (NGT 1.0).

2.2.1 Utvikling av NGT 1.0.

NGT 1.0 skal *baseres på de eksisterende emnesystemer i UBO og NB, med Humord som utgangspunkt*, og utvikles ved å *integrere* de nevnte emneordssystemene. NGT 1.0 skal ha termer på norsk bokmål for alle begreper, ellers ingen krav til språk.

Hovedmetodikk

Foreslått framgangsmåte er beskrevet i 7, og kan grovt oppsummeres som følger:

- Hvert vokabular, inkludert Humord, preprosesseres så godt det lar seg gjøre med automatiske metoder, dvs. strukturere/danne relasjoner, fjerne geografiske navn og sjangre, etc.
- Humord etableres som starttesaurus.
- Hvert av de andre vokabularene innlemmes etter tur i starttesaurusen, i denne rekkefølgen: Juridiske emneord, Realfagstermer, NBs emneordslister.

Håndteringen av begreper innenfor helsefag og psykologi skal utredes spesielt, med tanke på at MeSH er utbredt system i norske fagbibliotek, og er delvis oversatt til norsk.

Utviklingen av NGT 1.0 er estimert til å ta 2.5 år. Estimatet er heftet med en viss usikkerhet, se 7.1 (tidsplan).

Prosjektorganisasjon

Det anbefales å etablere en prosjekt- og koordineringsgruppe med 3 faste medlemmer: Prosjektleder (NB), utvikler/teknisk koordinator (NB) og tesaurusekspert (UBO). Disse bør jobbe full tid på prosjektet eller ha det som hovedaktivitet. I tillegg utpekes en gruppe på 10-15 bibliotekansatte (fra NB og UBO) som kan inndeles i midlertidige, fleksible arbeidsgrupper etter behov. Arbeidsgruppeledere deltar i prosjektgruppen i arbeidsgruppen virkeperiode. Prosjektet bør ha en styringsgruppe hovedsakelig bestående av KORG-faglig ledelse ved NB og UBO og ledet av NB. Se kapittel 8 for nærmere beskrivelse.

Tesaurussystem

Vi anbefaler å ta i bruk nyeste versjon av VocBench og representere begrepene i SKOS-XL (med noen potensielle utvidelser). VocBench ble prøvd ut i utviklingen av piloten (NGT 0.1). Se kapittel 9 for mer informasjon om valg av tesaurussystem og Appendiks 3 om utvikling av piloten.

Finansiering

NB bør finansiere/ansette prosjektleder og utvikler/teknisk koordinator for 3 år, samt infrastruktur. I tillegg foreslås å delfinansiere UBOs arbeid ved å kanalisere utviklingsmidler inn mot prosjektet, med forutsetning om en viss egenandel. Se for øvrig kapittel 10.

Publisering

NGT 1.0 bør gjøres tilgjengelig både via sluttbrukerverktøy for oppslag i selve tesaurusen og som nedlastbare, åpne, lenkede datasett, lisensiert i tråd med prinsippene om åpne data.

2.2.2 Drift av NGT

Vi anbefaler at NB er eier av og hovedansvarlig for Norsk generell tesaurus. Den første driftsorganisasjonen etter lansering av NGT 1.0 bør derfor være sentrert i NB med UBO og minst en institusjon fra Humord-samarbeidet som partnere.

Vi forslår å danne en redaksjonsgruppe med medlemmer fra nevnte organisasjoner, som koordinerer arbeidet med vedlikehold og videreutvikling. Ansvaret for den faglige kvaliteten i de ulike delene av NGT distribueres til grupper av fagreferenter/forskningsbibliotekarer ute i miljøene. Se kapittel 11 for nærmere beskrivelse.

Finansiering

NB bør dekke alle kostnader til infrastruktur og sekretariatsfunksjon. De andre deltakerinstitusjonene bør selv bekoste egen innsats.

2.2.3 Videre utvikling

Hvordan NGT skal utvikles etter versjon 1.0 ut over løpende vedlikehold er avhengig av mange forhold, og de fleste forslagene om dette kan derfor ikke være særlig konkrete. Følgende forslag står likevel fast:

- Det bør etableres mappinger fra NGT til Dewey. Dette blir hovedstrategien for å koble NGT til det internasjonale nettverket av emneautoriteter.
- NGT bør utvides språklig ved å oversette hele tesaurusen til engelsk og nynorsk. Samiske og kvenske termer bør innlemmes i den grad slike eksisterer for de ulike begrepene
- Større endringer når det gjelder faglig nedslagsfelt vil ofte ha sitt utspring i andre emnesystemer, eksempelvis lokalt forvaltede emneordslister som kan fylle ut mangler i NGT hvis de innlemmes i sistnevnte. Vi anbefaler konkret å vurdere TEKORD[1] (ingeniørfag) og Norsk idrettstesaurus[2] som kandidater for integrering.

Les mer om videre utvikling av NGT i kapittel 12.

[1] http://datahub.io/dataset/tekord
[2] http://www.nih.no/Documents/Bibliotek/Idrettstesaurus.pdf

2.3 Begrensninger i forprosjektet

Merk at Tesaurus forprosjekt kun har fokusert på NGT som ressurs og utvikling av denne. Arbeid knyttet til *innføring av NGT* i brukerinstitusjonene (i første omgang NB og UBO) er ikke planlagt eller utredet som del av Tesaurus forprosjekt. Innføringsaktiviteter er i utgangspunktet institusjonsspesifikke, og må planlegges separat i NB og UBO når NGT 1.0 nærmer seg ferdigstillelse, med samarbeid der det er naturlig og mulig.

3 Bakgrunn og motivasjon

Utgangspunktet for inneværende samarbeid er NBs og UBOs behov for emnemessig beskrivelse av samlingene sine. Begge institusjoner har i de senere år hver for seg analysert situasjonen på dette området med sikte på å utvikle en mer gjennomtenkt og helhetlig emnebeskrivelse av samlingene i de to bibliotekene.

3.1 Nasjonalbiblioteket

NB har siden 2011 hatt en intern aktivitet kalt *Emner i NB*, hvor oppgaven er å anbefale god praksis for emnebeskrivelse for alle materialtyper i NB. For materiale med verbalt innhold (dvs. ikke bildemateriale) kan prosjektets konklusjoner fra 2013 vedrørende emneord (Ohren, Rydland et al. 2013) oppsummeres som følger:

- Nåværende status er en svært ukoordinert bruk av emneord, med mange løsrevne, samlingsspesifikke emneordslister, - de aller fleste uten noen form for intern struktur, definisjoner eller synonymkontroll (Ohren, Rydland et al. 2012).
- Det anbefales å ta i bruk et allment, nasjonalt emneordssystem på tesaurusform for de aller fleste samlingene hvor emneord skal brukes. Emnesystemet bør bare inneholde innholdsbeskrivende emneord, ikke geografiske navn eller form/sjanger.
    - Et slikt emneordssystem finnes ikke pr i dag, men utprøving har indikert at tesaurusen Humord vil danne en god basis for videre utvikling.
    - Den resulterende tesaurusen vil fungere som en nasjonal ressurs og bør utvikles og forvaltes av flere større aktører, blant andre NB og den eksisterende Humord-redaksjonen.
- For spesialiserte fag og materialtyper hvor det finnes etablerte emnesystemer anbefales å bruke disse. I praksis gjelder dette:
    - MeSH[1] (norsk versjon), for helsefaglige tidsskriftartikler
    - FIAF[2] for bøkene i samlingen Trykte monografier om film

[1] Medical Subject Headings: http://www.nlm.nih.gov/mesh/
[2] International Federation of Film Archives: http://www.fiafnet.org/uk/

3.2 Universitetsbiblioteket i Oslo

Også ved UBO har det gjennom årene utviklet seg en nokså uensartet praksis for emnebeskrivelse, noe som klart framgikk av en analyse som ble utført i 2010 (Hegna, Almo et al. 2012):

Kartleggingen viser at UBO i dag registrerer emnedata i tilsammen 12 ulike MARC-felt i BIBSYS. I tillegg kommer de obligatoriske MARC-feltene 610-630 som gjelder emneinnførsler for personer, korporasjoner og titler. (s. 4)

…

Biblioteket nedlegger mye arbeid i emneordsindeksering, men manglende samordning vanskeliggjør gjenbruk av registreringsdata og skaper gjenfinningsproblemer. Sammen med mangelfull funksjonalitet i søkeprogrammene gjør dette at brukeren får relativt lite igjen for bibliotekets arbeidsinnsats. (s. 6)

På tross av uensartet praksis i UBO *som helhet* har man på enkelte avdelingsbibliotek lagt ned mye arbeid i å koordinere emneordsarbeidet. Humord startet opp i 1994 som en tesaurus innen humaniora. Etter hvert har tesaurusen blitt utvidet til også å inkludere samfunnsfag, men dette feltet er ennå under oppbygging. Humord vedlikeholdes i dag av en koordineringsgruppe under ledelse av en Humord-koordinator fra UBO/HumSam-biblioteket og med deltakere fra humanioraavdelingene ved universitetene i Bergen, Trondheim og Tromsø.

Også ved Realfagsbiblioteket har det i de senere årene vært gjort mye koordineringsarbeid på emneordssiden. Realfagstermer er i dag et kontrollert, pre-koordinert emneordsvokabular som i hovedsak dekker naturvitenskap, matematikk og informatikk. Realfagstermer består av ca. 15 000 frittstående emneord og rundt 16 000 emneord i streng.

Analysen fra 2010 anbefalte UBO å styre mot følgende mål:

- Få på plass en norsk tesaurus (Humord) som omfatter de fleste av UBOs fagområder
- Ta i bruk MeSH (pågående norsk oversettelse) for fagområdet medisin

3.3 Tesaurus forprosjekt – et samarbeid mellom UBO og NB

Ovenstående viser at UBO og NB har felles interesser i å få på plass en generell emneordsressurs, noe som danner et godt grunnlag for samarbeid. Med dette som utgangspunkt ble samarbeidsprosjektet *Tesaurus forprosjekt*[1] startet i mars 2014, med prosjektbeskrivelse utarbeidet av NB som grunnlag (se Appendiks 1), og hensikt og mål som beskrevet i 1.

[1] http://www.nb.no/Bibliotekutvikling/Kunnskapsorganisering/Tesaurus-forprosjekt

UBOs deltakelse i forprosjektet utgjør den ene halvparten av utviklingsprosjektet *På vei mot en generell, norsk tesaurus,* og er således finansiert ved utviklingsmidler fra 2014-tildelingen.

4 Innledende arbeid

På forprosjektets oppstartmøte 5. mars 2014 ble det bestemt å avholde et seminar med praksisfeltet så tidlig som mulig for å få inn synspunkter på behovet for og realismen i å utvikle en generell tesaurus i Norge.

På dette seminaret ble det etterlyst en oversikt over hvilke emneordssystem som er i bruk i bibliotekene. En undersøkelse om dette ble derfor gjennomført i juni 2014.

Prosjektgruppen ble også rådet til å undersøke situasjonen i andre land, blant annet hvilken rolle og ansvar andre lands nasjonalbibliotek har når det gjelder forvaltning av emneautoriteter.

I resten av dette kapittelet gjøres kort rede for arbeidet knyttet til disse tre temaene samt konklusjonene som er trukket på dette grunnlag.

4.1 Seminar med praksisfeltet

Seminaret ble avholdt i NBs lokaler i Oslo 30. april 2014. Ca. 90 deltakere fra over 30 institusjoner – alt overveiende fag- og forskningsbibliotek - deltok på seminaret. Programmet og presentasjonene ble publisert på prosjektsidene[1]. Det ble også laget en egen oppsummering av gruppearbeidet[2].

Fra arbeidet i gruppene og den påfølgende diskusjon, kan følgende framheves:

Alle gruppene uttrykte en positiv innstilling til en fellesløsning på emneordsfeltet. Det ble imidlertid klart påpekt at å utvikle en generell, nasjonal tesaurus vil være svært ressurskrevende. Det framkom også at det var delt syn på hva som ligger i begrepet "generell, nasjonal tesaurus". Hvor overordnet/dyp skal en slik tesaurus være? Et par grupper problematiserte det ulike behovet for granularitet som vil kreves når både bibliotek med generelle samlinger og bibliotek med høyst spesialiserte samlinger skal få dekket sine behov. Syv av elleve grupper så for seg en løsning der det utvikles en generell tesaurus som fungerer som en node/et nav/en overbygging i et nettverk av eksisterende tesauri – norske og flernasjonale – som man kan samsøke i/er mappet til hverandre.

Alternativer til å utvikle Humord til å bli en generell, nasjonal tesaurus vil være å oversette eksisterende, internasjonale vokabular til norsk. Library of Congress

[1] http://www.nb.no/Bibliotekutvikling/Kunnskapsorganisering/Tesaurus-forprosjekt/Paa-veg-mot-en-generell-nasjonal-tesaurus-Seminar
[2] http://www.nb.no/content/download/8091/80393/file/Tesaurus-seminar-gruppearbeid-oppsummering.pdf

Subject Headings (LCSH) og FAST (Faceted Application of Subject Terminology) ble nevnt.

Det ble sett på som naturlig at NB er koordinator i utviklingen av en eventuell generell tesaurus. Ellers uttrykte deltakerne en vilje til å bidra inn i prosessen med spesialkompetanse på sine fagfagfelt og til å bistå med etablering og vedlikehold av delhierarkier. Flere grupper understreket at NB må være overordnet faglig ansvarlig og at dette ikke er et dugnadsarbeid. Det må bevilges midler til arbeidet.

Når det gjelder målgrupper for tesaurusen kom det fram at mange mente at også folkebibliotekene kan være interessert i å bruke en slik tesaurus i indekseringsarbeidet. Andre så for seg at det er de store allmenne bibliotekene, særlig i fag- og forskningsbiblioteksektoren som vil være den mest aktuelle målgruppen. Noen nevnte også skolebibliotek og institusjoner i ABM-sektoren. Andre gikk bredere ut og inkluderte eksempelvis medieinstitusjoner, SINTEF og utdanningsinstitusjoner.

Seminaret ga prosjektgruppa mange gode innspill som er tatt med i det videre arbeidet.

4.2 Kartlegging av bruk av emneordsystem i Norge

I tråd med anbefalingene fra seminaret 30. april ble det i perioden 28. mai til og med 11. juni gjennomført en spørreundersøkelse om emneord og emneordsystemer i norske bibliotek. Samtlige fag- og forskningsbibliotek, fylkesbibliotekene samt noen større folkebibliotek ble invitert til å svare på undersøkelsen. Svarprosenten ble 35%, hvorav 86% fra fag- og forskningsbibliotek og 14% fra folke- eller fylkesbibliotek. Dette kan virke som en laber respons, men inspeksjon av dataene viser at de fleste store høgskoler og universitetsbibliotek hadde svart, og det er først og fremst disse som står for volumet i bestand blant fag- og forskningsbibliotekene[1]. Det er derfor grunn til å tro at representativiteten for fag- og forskningsbibliotekene i undersøkelsen er bedre enn svarprosenten tilsier.

De viktigste poengene fra undersøkelsen:

- 52% av respondentene bruker kontrollerte emneord, 42% bruker ukontrollerte. Totalt sett bruker altså langt de fleste - hele 94% - en eller annen form for emneord.
- Når det gjelder *ukontrollerte* emneord er det nesten bare fagbibliotekene som bruker slike, og UH-bibliotekene relativt sett mest.
- Det er en klar grense mellom folkebiblioteksiden og fagbiblioteksiden når det gjelder hvilke *kontrollerte* emnesystemer som brukes. Biblioteksentralens

[1] Bibliotekstatistikken 2013 (se http://ssb.no/177616/fag-og-forskningsbibliotek) viser at UH-bibliotekenes totale bestand i antall bind var 11,34 millioner, mens tilsvarende for resten av fagbibliotekene (spesialbibliotekene og bibliotekene ved helseinstitusjoner) var 3,32 millioner. NBs samlinger er ikke inkludert i noen av disse tallene.

emneord (BIBBI Emner[1]) er et rent folkebibliotekfenomen: Folke- og fylkesbibliotekene bruker nesten bare BIBBI, samtidig som nesten ingen fagbibliotek bruker BIBBI.
    - Unntak fra denne regelen er *Emneord for musikk*, en tesaurus utviklet på initiativ fra Norsk musikkbibliotekforening. Denne brukes til en viss grad av begge «sidene»
- Ser vi på fag- og forskningsbibliotekene som bruker kontrollerte emneord, preges bildet av to ting:
    - Halvparten at disse bibliotekene (34 av 70) utvikler sitt eget interne emneordssystem, uten å samarbeide med andre. Til sammen spenner disse over et bredt fagfelt.
    - MeSH er dominerende blant de som bruker eksisterende, etablerte emnesystemer, - 29 av 70 bruker dette. I tillegg til MeSH er det noen få (8 av 70) som bruker LCSH (Library og Congress Subject Headings), og Emneord for musikk (5 av 70).

Mer informasjon om resultatene fra emneordsundersøkelsen finnes på prosjektets nettsider[2].

Selv om svarprosenten ikke var så høy som vi hadde håpet, har denne undersøkelsen gitt prosjektgruppen en betydningsfull innsikt i bruken av emneord og emneordsystemer i Norge.

Av undersøkelsens kommentarfelter leser vi en overveiende positiv innstilling til et initiativ for å utvikle fellesløsninger på området. Ikke minst tyder den relativt store andelen fagbibliotek med hjemmesnekrede systemer på at fellesløsninger i form av et nasjonalt emneordssystem vil være av verdi.

4.3 Generelle emnesystemer i andre land

Parallelt med arbeidet med spørreundersøkelsen har vi kartlagt indekseringspolitikken i andre land. Her har prosjektgruppa i stor grad tatt utgangspunkt i IFLA-publikasjonen *Guidelines for Subject Access in National Bibliographies* (Jahns, 2010) og supplert og ajourført disse opplysningene. Nedenfor gis en kort beskrivelse av emnesystemene, se Appendiks 8 for mer informasjon.

Library of Congress Subject Headings (LCSH) danner mønster for mange nasjonale emneordsystemer og er dominerende i engelsktalende land. LCSH er oversatt og brukt i en del spansktalende land, i Canada også dels i fransk oversettelse. I Sverige har man også oversatt deler av de totalt fem millioner emneautoritetene. Strukturmessig er dette prekoordinerte emneord i streng. Gjennom import av metadata på e-bøker har mange norske fagbibliotek store mengder av poster med LCSH.

[1] http://www.bibsent.no/index.php/bibbi-autoriteter
[2] http://www.nb.no/Bibliotekutvikling/Kunnskapsorganisering/Tesaurus-forprosjekt/Spoerreundersoekelse-om-bruk-av-emneord-og-emneordsystemer

FAST, som utgår fra LCSH og dels er fasettert, er utviklet ut fra behovet for et enklere (og mindre) emneordssystem enn LCSH. FAST er blant annet i bruk hos New Zealands nasjonalbibliotek og i enkelte bibliotek i Australia.

Svenska ämnesord er utviklet av Kungliga Biblioteket og brukes i svenske fagbibliotek og i nasjonalbiblioteket. Dette systemet er utviklet ved at emneord brukt i svenske fagbibliotek ble samlet og "vasket". Systemet ble tatt i bruk i 2000. Nye emneord blir kontinuerlig mappet til eksisterende LCSH-termer. Svenska ämnesord består av tre ulike emneordslister: Svenska ämnesord (SAO), Tesaurus för grafiskt materiale (TGM) og Barnämneord (Barn).

I Finland er allmenne tesauri og spesialvokabular gjort tilgjengelig gjennom Finto. Dette omfatter også vokabular som ikke stammer fra biblioteksektoren.

Det tyske Schlagwortnormdatei brukes i Tyskland, Sveits og Østerrike. Det er laget på grunnlag av ordtilfang i emnesystemer i tyske bibliotek.

I Frankrike brukes RAMEAU av det franske nasjonalbiblioteket og av mange fag- og folkebibliotek.

I Italia er det utviklet en generell tesaurus (Nuovo Soggettario).

I de landene vi har undersøkt er det med andre ord ulike systemer i bruk. Noen er (som RAMEAU) delvis basert på LCSH, andre er utarbeidet lokalt. Et generelt inntrykk er at nasjonalbibliotekene har en førende og koordinerende rolle. Systemene varierer i størrelse og det er ulike emneordsstrukturer: fra prekoordinerte emneord til ontologier (Finland) og tesauri (Italia).

Uansett historikk og utgangspunkt legges det stor vekt på at emneordsvokabularene skal legges til rette for den semantiske weben.

4.4 Konklusjoner fra innledende arbeid

Ut fra 4.1 og 4.2 mener vi det er riktig å konkludere med at en nasjonal fellesløsning for emneord vil være velkommen og av stor verdi, spesielt for fag- og forsknings-bibliotekene.

Når det gjelder emneordssystemets struktur, argumenterte *Emner i NB* (Ohren, Rydland et al. 2013) for postkoordinert framfor prekoordinert system på denne måten:

> *De siste årene har det vokst fram en erkjennelse av at prekoordinerte emnesystemer er svært vanskelig å holde ved like, og dessuten er lite brukervennlige. Ikke minst har LCSH vært gjenstand for slik kritikk. Reglene for sammensetninger er komplekse, setter (for) store krav til indekserer, noe som i sin tur fører til mye feil. Tilsvarende erfaring er gjort med SÄO. Den noe større uttrykkskraften man oppnår ved å tillegge rekkefølgen av emneordene mening, blir i mange tilfeller svært kostbar. Prekoordinerte*

emneord er heller ikke godt tilpasset søkemotorer, og slett ikke semantisk web. Da OCLC ved århundreskiftet var på jakt etter et verbalt emnesystem som var enkelt å bruke, og som kunne optimalisere bruk av teknologi for metadata i Dublin Core, valgte de å ta utgangspunkt i vokabularet til LCSH, men bryte opp emnestrengene slik at emneordene fordeles på åtte fasetter (emner, sted, tid, hendelse, person, korporasjon, verk og form/sjanger). Også SÄO, opprinnelig bygd etter LCSH-modellen, er i gang med et forenklingsprosjekt.

Alt tatt i betraktning, vil vi ikke anbefale å satse på en prekoordinert emneordsliste.(s. 16)

Denne argumentasjonen gjelder fortsatt. Publikasjoner som sammenligner pre- og postkoordinering er relativt enige når det gjelder hvilke fordeler og ulemper som hefter ved pre- og postkoordinerte systemer. En litteraturstudie av Šauperl (Sauperl 2009) påpeker at prekoordinerte systemer riktignok gir mer presis gjenfinning enn postkoordinerte. *Eksempel:* En bok beskrevet ved de to emnestrengene *Silverwork – Peru* og *Art objects – Germany* i et prekoordinert system, vil få de fire frittstående emneordene *Silverwork*, *Art objects*, *Peru* og *Germany* i et postkoordinert system. I sistnevnte tilfelle vil boka uten tvil havne på trefflisten til spørringen *Silverwork AND Germany*, noe som er upresist.

På den annen side er prekoordinerte systemer svært kompetanse- og ressurskrevende å vedlikeholde, fordi konkateneringen av emneord i sekvenser nødvendigvis krever et regelverk (syntaks), som alle må forstå og følge. En undersøkelse blant indekserere i Library of Congress (Library of Congress 2007) i 2006 viste at alle respondentene – inkludert ansatte med mange års erfaring - syntes det var «somewhat difficult» eller vanskeligere både å indeksere og lage emnesekvenser med LCSH.

Et annet aspekt er at det neppe er realistisk å kunne indeksere automatisk med prekoordinerte emneord, - automatisk indeksering med en tesaurus er derimot ingen utopi, spesielt ikke når det gjelder fulltekstressurser.

I lys av ovenstående framstår ikke bibliotekfeltets innspill om å oversette (helt eller delvis) det prekoordinerte emneordssystemet LCSH som en enkel og ressursbesparende løsning, verken på kort eller lang sikt.

Heller ikke BIBBI Emner, som er så å si enerådende på folkebiblioteksiden vil være en fullgod løsning for fag- og forskningsbibliotek, - både fordi det er prekoordinert og fordi terminologien ikke er basert på vitenskapelige samlinger.

Prosjektgruppen mener fortsatt at en tesaurusstruktur vil gjøre det lettere å uttrykke semantiske relasjoner mellom emneord og dermed lettere kunne tilpasses den semantisk web, selv om det går på bekostning av spesifisitet i indekseringen. Et viktig moment i denne diskusjonen er selvfølgelig at det allerede eksisterer en

omfattende starttesaurus, nemlig Humord. Merk dessuten at mapping av NGT mot Dewey som foreslått i kapittel 12 vil innebære en viss integrasjon både mot LCSH og BIBBI.

5 Norsk generell tesaurus: Avgrensing og omfang

I dette kapittelet gjøres rede for hva NGT skal inneholde (og ikke inneholde) når det gjelder fag/emneomfang, språkformer på termene og emnetyper. Til slutt (i 5.5) gis noen generelle betraktninger om hvordan NGT kan forholde seg til andre emnesystemer.

Det aller meste i dette kapitlet gjelder NGT generelt. Under faglig dekning og språk (5.1 og 5.3) er det imidlertid skilt mellom første versjon (NGT 1.0) og videre utvikling, mens det som står om emnetyper gjelder generelt.

5.1 Faglig dekning

Det langsiktige målet er at NGT skal dekke alle fagområder med tilstrekkelig dybde, og at den utvikles i takt med utviklingen av samlingene den blir brukt til å indeksere, både når det gjelder domene (fagområder som dekkes) og granularitet.

Dette underbygges av at NBs samlinger er bygget opp rundt pliktavlevering. Kulturstatistikken viser riktignok at det er en overvekt av humanistiske og samfunnsmessige fag (Jensen 2014) i det som publiseres av monografisk materiale i Norge. NB håndterer imidlertid mye annet materiale enn bøker, blant annet norske og nordiske tidsskriftartikler, og det totale emnespekteret er vidt. Tilsvarende har UBO store samlinger innenfor de fagene det forskes og undervises i, og disse representerer til sammen alle de klassiske vitenskapelige disiplinene. Så selv om vi bare skulle tenke på NB og UBO sine samlinger, må et emneordssystem for indeksering av disse ha et bredt faglig nedslagsfelt.

5.1.1 Første versjon (NGT 1.0) = en sammenslåing av eksisterende vokabularer

Et viktig poeng er at NGT ikke skal bygges opp fra bunnen av. Som beskrevet ovenfor gjøres et betydelig arbeid med emneord og emneordssystemer både ved UBO og NB, og det er viktig at mest mulig av dette gjenbrukes i NGT. Det er derfor en forutsetning at eksisterende interne vokabularer – i første omgang fra UBO og NB – inngår som byggesteiner eller begrepskilder for NGT. Dette vil i stor grad påvirke NGTs initielle omfang og faglige dekning.

Prosjektet *Emner i NB* konkluderte i 2013 med at det beste utgangspunktet for en generell tesaurus for NBs samlinger er Humord (Ohren, Rydland et al. 2013).

Etter prosessen med seminar og kartlegging av emneordsarbeidet både i Norge og i sammenliknbare land, mener inneværende prosjektgruppe at dette fremdeles er en riktig tilnærming. Videre er det ønskelig at NGT allerede fra første versjon er

fullstendig i den forstand at både NB og UBO kan ta den i bruk uten å måtte bruke nåværende lokale emneordssystemer parallelt.

Dette oppnås ved å la første versjon av NGT (NGT 1.0) være resultatet av å integrere Humord med øvrige emneordssystem i UBO samt med NBs emneordslister, se Appendiks 2 for en kort beskrivelse av disse. En prosjektplan for dette arbeidet er spesifisert i kapitlene 6-10.

Det følger av dette at den faglige dekningen til NGT 1.0 vil være basert på samlinger i NB og UBO samt i de samarbeidende Humord-bibliotekene. Selv om det forventes at NGT 1.0 på denne måten vil få en bred faglig dekning, er det helt klart områder som er dårlig dekket. Typisk vil det i alle de integrerte vokabularene være emner i utkanten av vokabularets fagkrets (randemner) som er tynnere representert enn emner innenfor kjernefagene (kjerneemner). En grov, manuell analyse av dekningsgrad i Humord og Realfagstermer bekrefter dette. Eksempelvis er *ingeniørfag* et randemne både for Humord og Realfagstermer, og følgelig ikke særlig detaljert utarbeidet.

5.1.2 Videre utvikling

Videre utvikling av NGTs faglige nedslagsfelt etter at NGT 1.0 er lansert, kan for eksempel innebære utvidelse på emneområder hvor NGT 1.0 er tynt dekket. Slik utvidelse kan skje på flere måter, blant annet gjennom integrasjon av relevante eksisterende vokabularer fra norske institusjoner. Det er heller ikke utenkelig at man velger å begrense NGTs dekning på et emneområde. Dette kan være aktuelt dersom et godt fagspesifikt emnesystem på et gitt emneområde får en solid posisjon blant norske brukerinstitusjoner. Se for øvrig 5.5 om NGTs forhold til andre systemer.

Det er ikke lagt noen detaljplan for utvikling av NGT etter 1.0, men et overordnet veikart er skissert i kapittel 12.

Noe av utviklingsarbeidet nevnt ovenfor vil kreve konsentrert innsats over en periode. I tillegg utvikles NGT også gjennom det løpende vedlikeholdet som koordineres av NGTs driftsorganisasjon. Se kapittel 11 for informasjon om foreslått driftsmodell. Slik vi ser det i dag, bør inkludering av nye begreper være basert på litteraturbelegg/bestand, også i fortsettelsen. Det betyr at faglig komplettering avhenger av samlingsutviklingen hos alle som til enhver tid bruker NGT.

5.2 Emnetyper i NGT

Vokabularene som skal inngå i NGT 1.0 inneholder hver for seg mange typer emneord, for eksempel både geografiske navn, form/sjanger, i noen tilfeller korporasjoner og verk, i tillegg til innholdsbeskrivende emneord.

Prosjektet *Emner i NB* anbefaler imidlertid å begrense en generell tesaurus til å inkludere innholdsbeskrivende emneord (Ohren, Rydland et al. 2013) og argumenterer spesielt for å utelukke geografiske steder og form/sjanger:

- Navn på geografiske steder forvaltes av egne instanser (Kartverket for navn i Norge). Stedsnavn som emneord bør forholde seg til disse og håndteres for seg, analogt til person- og korporasjonsautoriteter. Dette er aktualisert ved at norske kartdata nå er åpent tilgjengelig, noe som åpner helt nye muligheter for digitale tjenester med utgangspunkt i autoriserte stedsnavn.
- Når det gjelder form/sjanger, finnes det etablerte, internasjonale systemer for dette på flere detaljnivåer. Library of Congress Genre Form Terms[1] (LCGFT) er nylig oppdatert og realisert som tesauri, og disse bør prøves ut før man går inn for å lage et eget system.

5.2.1 Emnetyper som ikke skal inkluderes i NGT

Prosjektgruppen støtter tankegangen fra *Emner i NB* og anbefaler at form/sjanger og de aller fleste typer *navngitte entiteter* utelukkes fra NGT. Konkret gjelder dette:

- Navn på geografiske steder (se ovenfor).
- Form og sjanger (se ovenfor).
- Navn på (reelle) korporasjoner og personer: Disse hentes fra autoritetsregister for navn og skal ikke være del av en generell norsk tesaurus. Navngitte, fiktive skikkelser vil derimot kunne inkluderes (eksempel: James Bond).
- Verk: Omtalte verk vil ikke inngå i tesaurusen. Konstruksjonen av emneord for verk (forfatter/tittel eller standardtittel) følger gjeldende katalogiseringsregler og kodes i dedikerte MARC-felt. Foreløpig finnes ikke noe eget autoritetsregister over verk (men bør etableres på sikt).

5.2.2 Emnetyper som skal inkluderes i NGT

Følgende emnetyper skal inngå i NGT:

- Innholdsbeskrivende emneord: NTG vil inneholde alle typer emneord som beskriver dokumenters innhold (ikke ytre form). Dette inkluderer også *innholdsbeskrivende emneord av allmenn karakter* og *generelle emneord* som behandler dokumentets indre form (Hjortsæter 2009)
- Tidsperioder og tidsepoker er en del av Humord i dag og vil også være fasetter som videreføres i NGT.
- Navngitte entiteter av visse typer, for eksempel fiktive personer/skikkelser, religiøse skikkelser, gjenstander og kunstverker skal kunne inkluderes i NGT. Slike er også en del av Humord i dag.

5.3 Språkform

NGT skal være flerspråklig i den forstand at datamodell og infrastruktur må støtte at ethvert begrep i tesaurusen har betegnelser/termer på et vilkårlig antall språk. Hvilke språk som faktisk skal inkluderes er en annen sak. For NGT 1.0 er kravet at alle begreper skal ha betegnelser/termer på norsk bokmål. På sikt er det aktuelt at alle

[1] Se http://id.loc.gov/authorities/genreForms.html for søking og http://www.loc.gov/catdir/cpso/genreformgeneral.html for informasjon

offisielle språk i Norge samt engelsk er representert i NGT. Utvidelse til flere språk er foreslått i kapittel 12.

5.4 Brukergrupper, bruksområder og bruksrettigheter

Intensjonen er at NGT fra første versjon (NGT 1.0) av skal være åpen og tilgjengelig for alle på flere måter, - både via sluttbrukerverktøy for oppslag i selve tesaurusen og som nedlastbare, åpne, lenkede datasett.

Forutsatt at det ikke er juridiske hindringer for dette, *er NGT ment for enhver som kan ha nytte av den, uansett til hvilket bruk*, det være seg kommersielt eller ikke-kommersielt. NGT skal derfor lisensieres i tråd med prinsippene om åpne data.

Potensielle bruksområder for NGT omfatter blant annet:

- Indeksering av informasjonsressurser
- Som språklig/terminologisk verktøy
- Utvikling av tjenester med utgangspunkt i dataene, typisk
    - Integrering i tredjeparts systemer, for eksempel til
        - Søkehjelp (integrering i søkegrensesnitt)
        - Autoritetskontroll under katalogisering (integrering i biblioteksystem)
    - Nye tjenester basert på NGT-dataene, eventuelt kombinert med andre datasett

Som indekseringsredskap er NGTs *primære* målgruppe fag- og forskningsbibliotek, i hvert fall foreløpig (for NGT 1.0). Det er tidligere påpekt at tilfanget av begreper og termer i NGT alltid vil være påvirket av samlingene den brukes til å indeksere. NGT 1.0 har sitt utspring i en vitenskapelig tesaurus (Humord) og henter sitt øvrige innhold fra vitenskapelige bibliotek. Den vil dermed ha et relativt sterkt innslag av vitenskapelige terminologi i forhold til mer populære begreper. Som sådan vil NGT 1.0 trolig være best tilpasset bruk i fag- og forskningsbibliotek.

Når det er sagt, vet vi at emneordene brukt i NORART vil utgjøre et vesentlig innslag i NGT 1.0, og siden under 5% av artiklene i artikkeldatabasen NORART er karakterisert som vitenskapelige[1], er det grunn til å anta at en stor del av emneordene derfra vil ha en populær form. Vi må derfor regne med at NGT 1.0 vil ha et visst innslag av populær terminologi (f.eks. i form av henvisninger/synonymer), noe som kan gjøre den egnet til indeksering i andre institusjoner/samlinger, så som folkebibliotek, arkiver og museer.

5.5 Forholdet til andre emnesystemer

NGT skal virke i en verden der det allerede finnes mange andre emnesystemer, og det er viktig å være bevisst på hvordan NGT best mulig kan innrettes i forhold til

[1] Antall artikler i NORART oppgitt som vitenskapelige er 24582 av totalt 573229 artikler, dvs. 4,3% vitenskapelige.

eksisterende vokabularer. Vi ser for oss 3 hovedmåter NGT kan relateres til andre emnesystemer på:

1. *Integrere* andre emnesystemer inn i NGT: Dette betyr å *innlemme* emnesystemene, slik at de blir en del av NGT. Som det framgår i det foregående, blir NGT 1.0 i praksis en integrasjon av alle lokale vokabularer ved UBO og NB. Senere vil det også være aktuelt å integrere lokale emnesystemer fra andre institusjoner som er interessert i å delta i samarbeidet, spesielt på områder der vi anser at NGT bør styrkes.
2. *Tilpasse NGTs faglige dekning* til andre emnesystemer: På noen fagområder eksisterer det etablerte emneordsystemer som brukes av mange og store aktører. Da må det vurderes om, eventuelt hvordan, NGT skal representere emner innenfor disse områdene. Emneordsundersøkelsen viste at MeSH er et svært sentralt system på områdene medisin og helse samt psykologi. Disse områdene er også sterkt representert i noen av vokabularene som skal inngå i NGT 1.0. Vi ser det derfor nødvendig å avklare hvordan NGT skal forholde seg til dette allerede i første versjon. AGROVOC på området miljø og landbruk er en kandidat for tilsvarende vurdering på et senere stadium.
3. *Mappe* til andre emnesystemer: Noen få, viktige emnesystemer, har status som nav/hub-er i det internasjonale nettverket av emneautoriteter. DDC og LCSH (ev. FAST) er eksempler på slike. Måten å knytte NGT opp mot dette nettverket er å mappe NGT mot f.eks. Dewey Decimal Classification (DDC), ved å etablere lenker fra begreper i NGT til relaterte begreper i DDC.

6 Plan for utvikling av Norsk generell tesaurus 1.0

I dette kapittelet er det spesifisert en plan for utvikling av første versjon av NGT (NGT 1.0). Planen inneholder følgende deler:

- En kort beskrivelse av hva utviklingsprosjektet skal resultere i
- Aktivitets- og tidsplan: Hva vi må gjøre for å oppnå målet?
- Prosjektorganisasjon: Grupper, roller og kompetansebehov
- Andre ressursbehov, for eksempel tesaurussystem
- Forslag til finansiering

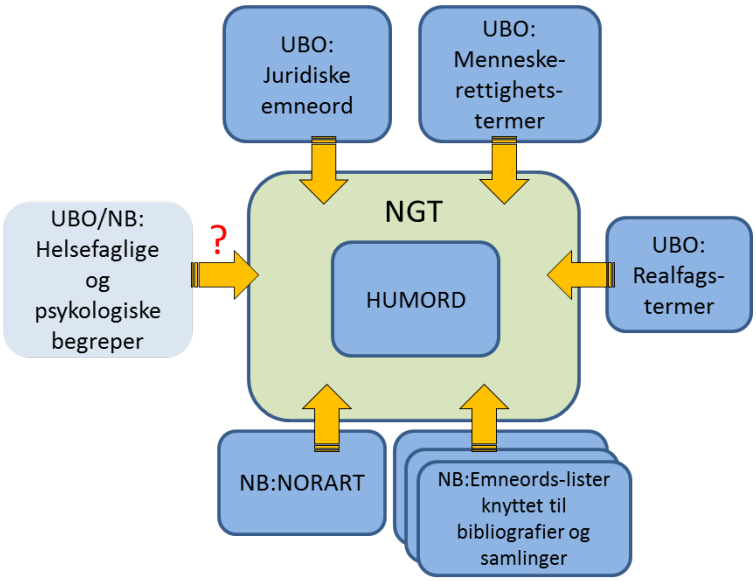
6.1 Mål

Målet med prosjektet er å etablere en tesaurus – NGT 1.0 – hvor de viktigste emneordssystemene fra UBO og NB er integrert i en felles struktur, og at alle begrepene har persistente identifikatorer i navnerommet data.nb.no. Følgende emnesystemer skal inn:

- Humord
- Juridiske emneord, inkludert Menneskerettighetstermene
- Realfagstermer

- Emneord fra NB (NORART, forfatter- og spesialbibliografier, samisk bibliografi)

Vokabularene er beskrevet i Appendiks 2. Humord, som det mest generelle, og det eneste som har reell tesaurusform, danner utgangspunktet for integreringsprosessen. Se Figur 1 for anskueliggjøring.



Figur 1 "I begynnelsen var Humord" – NGT fra Humord til en helhetlig, flerfaglig tesaurus.

Som illustrert i figuren er det en usikkerhet knyttet til håndteringen av helsefaglige og psykologiske begreper (se også 5.2 punkt 2). Dette må avklares og beslutning gjennomføres innenfor rammen av NGT 1.0.

NGT 1.0 skal ha termer på norsk bokmål for alle begrepene. Datamodellen må støtte generell flerspråklighet. Eventuelle eksisterende termer i andre språk fra kildevokabularene inkluderes i NGT 1.0.

NGT 1.0 skal i det alt vesentlige inneholde bare *innholdsbeskrivende emneord*. I tillegg inkluderes tidsepoker og navngitte fiktive personer og skikkelser, i den grad slike finnes i kildevokabularene. Dette er også tenkt å gjelde senere versjoner.

Form/sjanger og navngitte entiteter som geografiske steder, korporasjoner og personer skal ikke inkluderes i NGT 1.0, ei heller i senere versjoner.

Intensjonen er å gjøre NGT 1.0 åpent tilgjengelig for enhver type anvendelse, dvs. at dataene tilordnes en mest mulig åpen lisens.

7 Utvikling av NGT 1.0 – Aktiviteter og tidsplan

Aktivitetene som må gjennomføres for å nå målet er beskrevet i følgende tabell. Merk at det ikke ligger en garantert rekkefølge i nummereringen av aktivitetene. Arbeidsflyten er visualisert i Figur 2.

Aktivitet	*Beskrivelse*
1	**Etablere prosjektet** **Ansvarlig:** Prosjekteiere, dvs. NB og UBO Det nedsettes en styringsgruppe, som deretter får det formelle ansvaret for prosjektet. Den formelle etableringen av prosjektet består i å inngå de nødvendige avtaler mellom prosjekteierne og sette opp prosjektorganisasjonen.
1.1	**Formalisere prosjektet** Prosjektet bør formaliseres ved en skriftlig avtale mellom NB og UBO, hvor det framgår hva hver av partene skal bidra med.

1.2	**Danne prosjektorganisasjon**
	
	Ansvarlig: Styringsgruppen
	
	Prosjektorganisasjonen er nærmere beskrevet i kapittel 8. Etablering av denne består i å:
	
	• Ansette prosjektleder (3 års ansettelse)
	• Leie inn eller ansette utvikler/teknisk koordinator
	• Identifisere tesauruskoordinator
	• Etablere prosjekt- og koordineringsgruppe
	• Identifisere fageksperter i NB og UBO
	  o KORG-faglige, til å delta i de oppgavebaserte arbeidsgruppene
	  o Domeneeksperter

2	**Etablere teknisk infrastruktur og systemstøtte** **Ansvarlig:** Prosjekt- og koordineringsgruppe, hovedsakelig utvikler/teknisk koordinator Infrastrukturen for NGT skal installeres på NB, og må derfor avklares med driftsansvarlig enhet i NB.
2.1	**Anskaffe og implementere tesaurussystem** **Ansvarlig:** Prosjekt- og koordineringsgruppen, hovedsakelig utvikler/teknisk koordinator Se på anbefalingene fra forprosjektet, og sjekk at faktaene de er basert på fortsatt gjelder. Ta i betraktning eventuelle nye momenter (f.eks. nye behov, nye markedsaktører, ny utvikling av de vurderte produktene), velg og anskaff tesaurussystem. Merk at arbeidet med dette trolig vil øke med avstanden i tid mellom forprosjekt og hovedprosjekt.

Aktivitet	*Beskrivelse*
2.2	**Kvalitetssikre representasjonsmodell** **Ansvarlig:** Prosjekt- og koordineringsgruppe, hovedsakelig utvikler/teknisk koordinator. I forprosjektet ble det utarbeidet et forslag til hvordan NGT og begrepene den består av bør representeres. Modellen bør kvalitetssikres og eventuelt oppdateres. Spesielt bør det vurderes om muligheten for å operere med mikrotesauri skal være med i modellen.
2.3	**Anskaffe eller utvikle støtte for integrering av vokabularer** **Ansvarlig:** Prosjekt- og koordineringsgruppe, hovedsakelig utvikler/teknisk koordinator. Det er behov for systemstøtte til selve integreringen mellom vokabularer. I motsetning til tesaurussystemer som sådan, er dette er i liten grad hyllevare. Det er derfor ikke utredet spesielt i forprosjektet, selv om noen av de vurderte tesaurussystemene har støtte for såkalt «vocabulary alignment» eller «vocabulary merging» , - eller har dette på planen. Spesielt bør det sees på om mappingverktøyet utviklet i UBOs prosjekter kan tilpasses for vårt bruk. Støttefunksjonene/-verktøyene vi skaffer på dette området kalles *integreringsverktøy* i resten av dette dokumentet.
2.4	**Anskaffe publiseringsplattform for data** **Ansvarlig**: Prosjekt- og koordineringsgruppe Dataene skal lagres, vedlikeholdes og være søkbare i tesaurussystemet, men det er også ønskelig å publisere datasett som kan lastes ned, og som er daterte/versjonerte og dessuten beskrevet med metadata. Det er også ønskelig å sette opp regelmessig, automatisk høsting fra tesaurussystemet til dataplattformen.
3	**Preprosessering av de enkelte vokabularer** De ulike vokabularene som skal integreres har ulik struktur og detaljrikdom, og er brukt på ulike samlinger. Se nærmere beskrivelse i Appendiks 2. Før forsøk på integrering gjøres, bør hvert enkelt vokabular preprosesseres slik at mest mulig av informasjonen som er innbakt i dets originale kontekst hentes ut og utnyttes til å strukturere og berike vokabularet. Merk at den konkrete bearbeidelsen beskrevet for hvert vokabular nedenfor er basert på innsikten vi har under planleggingen. Selve prosesseringen vil i seg selv kunne avdekke forhold som tilsier *mer*

Aktivitet	*Beskrivelse*
	eller *annen* behandling enn foreslått i de neste delkapitlene.
3.1	**Preprosessere Humord - etablere NGTs starttesaurus NGT_0** **Ansvarlig:** Arbeidsgruppe for prosessering av Humord **Andre deltakere:** Utvikler Humord er nærmere beskrevet i Appendiks 2. Preprosesseringen av Humord innebærer å utføre følgende: • Utelate alle geografiske navn, betegnelser på form, navn på (reelle) personer og korporasjoner og annet som ikke hører med i NGT. • Konvertere resultatet til vedtatt representasjonsmodell o Inkluder informasjon om proveniens/kilde for alle begreper • Implementere modellen i anskaffet tesaurussystem • Identifisere og isolere medisinske/helsefaglige termer • Omstrukturere i henhold til ny toppstruktur (se 0) • Rydde i nåværende hierarkier etter behov Rekkefølgen ovenfor er basert på en antakelse om at det er lettest å gjøre omstrukturering ved hjelp av tesaurussystemets redigeringsfunksjoner. Hvis det derimot viser seg at dette kan utføres ved relativt enkle skript, kan det vise seg mer effektivt å utvikle og kjøre disse før vokabularet importeres i tesaurussystemet. Resultatet av aktiviteten blir «baseline NGT», heretter kalt NGT_0, som de andre vokabularene suksessivt skal integreres med, og som således danner grunnsteinen for NGT 1.0.

3.2	**Preprosessere juridiske og menneskerettslige emneord**
	
	Ansvarlig: Arbeidsgruppe for prosessering av juridiske og menneskerettslige emneord
	
	Andre deltakere: Utvikler/teknisk koordinator
	
	Juridiske og menneskerettslige termer er nærmere beskrevet i Appendiks 2. Preprosesseringen av disse termene innebærer å utføre følgende:
	
	• Prosessere menneskerettighetstermene:
	  o Utelate eventuelle termer av typer som ikke skal med i NGT (hovedsakelig termer for sjanger/form og geografiske navn)
	  o Oversette termene til norsk. Hvis mulig/hensiktsmessig bør dette gjøres automatisk (f.eks. ved hjelp av Google translate) med etterfølgende manuell revisjon
	• Prosessere juridiske emneord (emneordene knyttet til L-

Aktivitet	*Beskrivelse*
	skjema): o Løse opp emnestrenger o Utelate eventuelle termer av typer som ikke skal med i NGT (hovedsakelig termer for sjanger/form og geografiske navn) • Finne overlapp (eksakt leksikalsk match) mellom menneskerettighetstermene (på norsk) og juridiske emneord • Konvertere den samlede mengde ulike termer til vedtatt representasjonsmodell • Inkludere informasjon om proveniens/kilde for alle begreper. For termene som utgjør overlappet angis begge kildene • Ta vare på knytningen til L-skjemanummer (klassifikasjonskode) • Identifisere emnegrupper og/eller hierarkiske strukturer. Knytningen til L-skjema kan muligens brukes til dette o Identifiser og isoler medisinske/helsefaglige termer, i den grad det er mulig[1].
3.3	**Preprosessere Realfagstermer** **Ansvarlig:** Arbeidsgruppe for Realfagstermer **Andre deltakere:** Utvikler/teknisk koordinator Realfagstermer er nærmere beskrevet i Appendiks 2. Preprosesseringen av disse termene innebærer å utføre følgende: • Løse opp emnestrenger • Konvertere termene til vedtatt representasjonsmodell o Inkluder informasjon om proveniens/kilde for alle begreper • Utelate termer av typer som ikke skal med i NGT (termer for sjanger/form og geografiske navn) • Identifisere emnegrupper og/eller hierarkiske strukturer, hvis mulig. Eksempelvis, via metadata kan eierinstitutt til bøker beskrevet med et gitt emneord identifiseres, noe som representerer en viss faglig inndeling. • Identifisere og isolere medisinske/helsefaglige termer, i den grad det er mulig
3.4	**Preprosessere NBs emneordslister** **Ansvarlig:** Arbeidsgruppe for NB-emneord **Andre deltakere:** Utvikler/teknisk koordinator • Prosessere forfatterbibliografiene

Aktivitet	*Beskrivelse*
	o Lage en omforent emneordsliste av alle forfatterbibliografiene. • Prosessere emneordene for Samisk bibliografi o Løse opp emnestrenger o Utelate termer av typer som ikke skal med i NGT (termer for sjanger/form og geografiske navn) • Prosessere emneordene brukt i NORART På grunn av det store volumet (ca 40000 emneord) og det faktum at de i utgangspunktet ikke er relatert på noe vis, er det av stor betydning å strukturere disse så godt som mulig før de integreres med resten av NGT. Nøkkelen til dette ligger i metadataene, som er av god kvalitet og gir flere muligheter til statistisk analyse: o samforekomst mellom emneord med DDK-nummer o samforekomst mellom emneord og tidsskrift, samt sistnevntes DDK-nummer o samforekomst mellom emneord og ord i titler Ved å bruke Formal Concept Analysis[1] (FCA) bør det være realistisk å oppnå en viss gruppering/strukturering av emneordene, se Appendiks 7 for nærmere beskrivelse. I denne prosessen håper vi blant annet å kunne identifisere emneord innenfor helsefag og psykologi, geografiske navn og formbetegnelser. • Utelate termer for sjanger/form og geografiske navn, i den grad disse er identifisert av prosessen ovenfor. Manuell revisjon for å få med alle kan være nødvendig. • Utelate/isolere termer innenfor helsefag og psykologi, i den grad disse er identifisert av prosessen ovenfor. Manuell revisjon for å få med alle kan være nødvendig. • Finne parvis overlapp (eksakt leksikalsk match) mellom alle tre vokabularene • Konvertere den samlede mengde ulike termer til vedtatt representasjonsmodell • Inkludere informasjon om proveniens/kilde for alle begreper. For termene som er inkludert i to eller alle tre vokabularer, angis begge/alle kildene
4	**Avklare håndtering av begreper innenfor helsefag og psykologi i eksisterende vokabularer** Alle vokabularene som skal integreres - med sannsynlig unntak av

[1] En statistisk-matematisk metode for å utlede et begrepshierarki fra en samling objekter (her emneord) og deres egenskaper

Aktivitet	*Beskrivelse*
	Juridiske emneord - inneholder en betydelig mengde begreper innenfor medisin/helsefag og psykologi. Samtidig er store deler av MeSH[1] – de facto internasjonal standard for medisinske emneord - oversatt til norsk[2], noe som gjør det nødvendig å se spesielt på hvordan NGT skal forholde seg til dette.
4.1	**Utrede håndtering av begreper innenfor helsefag og psykologi i eksisterende vokabularer** **Ansvarlig:** Arbeidsgruppe for vokabular innen helsefag og psykologi Problemstillingen beskrevet ovenfor utredes. Tre hovedtilnærminger peker seg ut: • Fjerne medisinske og eventuelt psykologiske begreper fra NGT. MeSH anbefales til indeksering på disse feltene. • Beholde de eksisterende medisinske begreper i NGT, rydd etter behov, som i andre hierarkier, men ikke videreutvikle denne delen. Begrepene bør mappes til MeSH, og MeSH anbefales til indeksering. • Beholde de eksisterende medisinske begreper i NGT, rydd etter behov, som i andre hierarkier, og videreutvikle vokabularet som et tilbud til brukere som ønsker et mer overordnet og generelt medisinsk vokabular enn det MeSH tilbyr. Vokabularet bør mappes til MeSH. Alternativene utredes og løsning anbefales.
4.2	**Bestemme hvordan begreper innenfor helsefag og psykologi skal håndteres i NGT** **Ansvarlig:** Styringsgruppen Anbefalingen fra Akt. 4.1 behandles og vedtak gjøres
5	**Utføre vedtak om begreper innenfor helsefag og psykologi** **Ansvarlig:** Arbeidsgruppe medisinsk vokabular Dersom konklusjonen fra Akt. 4.2 (se ovenfor) innebærer at NGT skal inneholde medisinske begreper, må den nødvendige integrasjonen utføres innenfor rammen av versjon 1.0. Eventuelt vedtak om mapping til MeSH kan om nødvendig vente til etter lansering.
6	**Utarbeide toppstruktur for NGT** **Ansvarlig:** Prosjekt- og koordineringsgruppe

[1] http://www.nlm.nih.gov/mesh/

[2] http://www.kunnskapssenteret.no/verktoy/medisinske-og-helsefaglige-termer-pa-norsk-og-engelsk

Aktivitet	*Beskrivelse*
	Andre deltakere: Potensielt alle som bemanner arbeidsgruppene. Resultatet av denne aktiviteten skal være en god toppstruktur for NGT. Dagens Humord har som kjent et bredere faglig omfang enn humaniora og samfunnsfag, i og med at det finnes hierarkier for et begrenset antall emner både innenfor naturvitenskapene, rettsvitenskap og medisin. Når nå spesialvokabularer på disse områdene skal integreres, er det naturlig av forvalterne av disse blir å betrakte som autoriteter for strukturering av disse områdene. Gjennom en bred diskusjon på tvers av alle involverte fag/emner skal det utarbeides en omforent toppstruktur for NGT. Her er det naturlig å ta utgangspunkt i dagens Humord, samtidig som det må tas hensyn til de nye fagområdene som skal inn. Det bør utarbeides et initielt forslag så tidlig som mulig, som så revideres når/hvis integreringen av vokabularene avdekker behov for dette.
7	**Integrere de preprosesserte vokabularene med NGT** Denne hovedaktiviteten innebærer å inkorporere de preprosesserte vokabularene i NGT etter tur. **Overordnet tilnærming** Det mest fornuftige er antakelig å integrere ett vokabular om gangen, slik at vi til enhver tid har en fungerende tesaurus som danner en god struktur å integrere et nytt vokabular inn i. En slik tilnærming vil også resultere i en metode/prosess som kan gjenbrukes når/hvis eksterne vokabularer skal integreres på et senere stadium. For hvert vokabular V som skal integreres med eksisterende versjon av NGT er det viktig å tidligst mulig identifisere begreper som allerede finnes fra før. Hvis V er en flat liste uten henvisninger, har de ulike termene ingen kontekst ut over det vi vet om samlingen(e) de er brukt på. Vi antar derfor at alle begreper c i V hvor foretrukket term er lik foretrukket term i NGT kan betraktes som allerede integrert i NGT. Det bør imidlertid knyttes noter til begrepet i NGT, som angir hvilke vokabular(er) det kommer fra. Rekkefølgen vokabularene integreres i er valgt ut fra en tanke om at vokabularene med en sterk faglig forankring bør inn før de som ikke i samme grad er kvalitetssikret av fagspesialister. Eksempelvis antar vi at termene i Juridiske emneord i større grad er kvalitetssikret av fagspesialister enn eventuelle juridiske begrep i NORARTs emneord, og at de derfor vil utgjøre et bedre utgangspunkt for jus-delen av NGT. Integreringen vil forhåpentligvis kunne utføres ved et interaktivt samspill mellom system og menneske. Dette avhenger av graden av systemstøtte, dvs. hvor gode integreringsverktøy vi oppnår som resultat

Aktivitet	*Beskrivelse*
	av aktivitet 2.3. For hvert vokabular som integreres, skal de berørte delene av NGT kvalitetssikres i samarbeid med fageksperter. Dette inngår i alle aktivitetene 7.1-7.3.
7.1	**Integrere Juridiske og menneskerettslige emneord inn i NGT_0** **Ansvarlig:** Arbeidsgruppe for juridiske emneord, Arbeidsgrupper for berørte NGT-hierarkier. Se generell beskrivelse i hovedaktivitet (7) ovenfor. Hovedstegene i integreringsprosessen er som følger: • Identifisere match mellom det juridiske vokabularet og NGT_0, i henhold til mulighetene i integreringsverktøyet o Minimumsvarianten vil være å identifisere direkte overlapp med NGT_0 basert på eksakt leksikalsk match o Kontrollér og revidér matchene identifisert av integreringsverktøyet. Eventuelle tvilstilfeller markeres spesielt • Integrere manuelt alle begrep i det juridiske vokabularet hvor det ikke ble funnet en match med NGT_0. Eventuelle spørsmål/tvil markeres spesielt • Tvilstilfellene avklares sammen med andre i arbeidsgruppa, samt eventuelle fagspesialister • Kvalitetssikring: Det integrerte vokabularet kvalitetssikres i regi av Prosjekt- og koordineringsgruppen Resultatet av ovenstående er NGT_1 som består av NGT_0 samt juridiske og menneskerettslige emneord
7.2	**Integrere Realfagstermer inn i NGT** **Ansvarlig:** Arbeidsgruppe for realfagstermer, Arbeidsgrupper for berørte NGT-hierarkier. Se generell beskrivelse i hovedaktivitet (7) ovenfor. Hovedstegene i integreringsprosessen er som følger: • Identifisere match mellom Realfagstermer og NGT_1, i henhold til mulighetene i integreringsverktøyet o Minimumsvarianten vil være å identifisere direkte overlapp med NGT_1 basert på eksakt leksikalsk match o Kontrollere og revidere matchene identifisert av integreringsverktøyet. Eventuelle tvilstilfeller markeres spesielt • Integrere manuelt alle begrep i Realfagstermer hvor det ikke ble funnet en match med NGT_1. Eventuelle spørsmål/tvil markeres spesielt • Tvilstilfellene avklares sammen med andre i arbeidsgruppa,

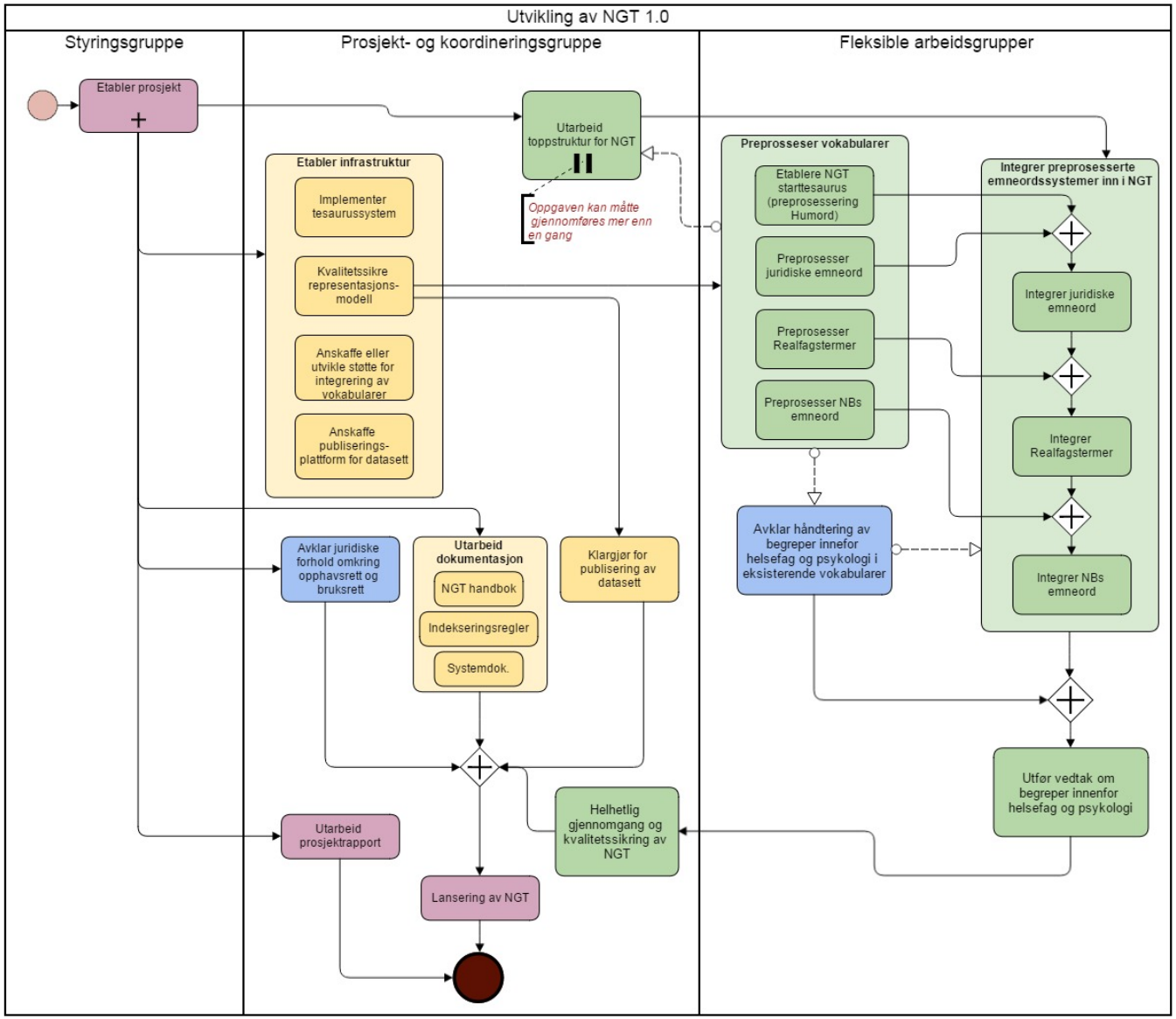
Aktivitet	*Beskrivelse*
	samt eventuelle fagspesialister • Kvalitetssikring: Det integrerte vokabularet kvalitetssikres i regi av Prosjekt- og koordineringsgruppen Resultatet av ovenstående er NGT_2 som består av NGT_1 samt Realfagstermer
7.3	**Integrere NBs emneordslister inn i NGT** **Ansvarlig:** Arbeidsgruppe for NB-emneord, Arbeidsgrupper for berørte NGT-hierarkier. Se generell beskrivelse i hovedaktivitet (7) ovenfor. Hovedstegene i integreringsprosessen er som følger: • Identifisere match mellom NBs emneord og NGT_2, i henhold til mulighetene i integreringsverktøyet o Minimumsvarianten vil være å identifisere direkte overlapp med NGT_2 basert på eksakt leksikalsk match o Kontrollere og revidere matchene identifisert av integreringsverktøyet. Eventuelle tvilstilfeller markeres spesielt • Integrere manuelt alle begrep i NBs emneord hvor det ikke ble funnet en match med NGT_2. Eventuelle spørsmål/tvil markeres spesielt • Tvilstilfellene avklares sammen med andre i arbeidsgruppa, samt eventuelle fagspesialister • Kvalitetssikring: Det integrerte vokabularet kvalitetssikres i regi av Prosjekt- og koordineringsgruppen Resultatet av ovenstående er NGT_3 som består av NGT_2 samt NBs emneord.
8	**Helhetlig gjennomgang og kvalitetssikring av NGT** **Ansvarlig:** Prosjekt- og koordineringsgruppen **Andre deltakere:** Arbeidsgrupper for berørte NGT-hierarkier Det skal gjøres en sluttgjennomgang og kvalitetssikring av hele NGT. Nødvendig detaljnivå på gjennomgangen vil avhenge av kvalitetssikringen utført i 7.1-7.3. Gjennomgang av hierarkiene bør gjøres av arbeidsgrupper sammensatt for dette formålet. Gruppene bør settes sammen slik at den enkelte i minst mulig grad må revidere eget arbeid.
9	**Klargjøring av NGT for «markedet»** Denne aktiviteten innebærer å gjøre det som skal til for at NGT kan tas i bruk av de som ønsker det.
9.1	**Avklare juridiske forhold omkring opphavsrett og bruksrett**

Aktivitet	*Beskrivelse*
	Ansvarlig: Prosjekt- og koordineringsgruppe Det må avklares hvem som skal ha opphavsretten til NGT, og eventuelle juridiske grep for å tilordne opphavsretten må utføres. Det må også bestemmes hvilken lisens som skal legges på dataene og eventuell medfølgende programkode. Det bør legges til grunn at lisensene skal tillate friest mulig bruk, både kommersiell og ikke-kommersiell. I dette arbeidet blir det nødvendig å hente inn juridisk kompetanse, i første omgang søke råd hos NBs jurist med opphavsrett som spesialområde.
9.2	**Klargjøre for publisering av datasett** **Ansvarlig:** Tesaurus- og koordineringsgruppe **Andre deltakere:** Utvikler/teknisk koordinator Aktiviteten innebærer å: • bestemme hvilke formater dataene skal publiseres i. Aktuelle formater er blant andre SKOS(-XL), MADS og MARC21 • utvikle skript som produserer de valgte formatene • bestemme hvilke metadata som bør følge datasettene • hvis mulig, konfigurere datapubliseringsplattformen til automatisk, regelmessig høsting fra tesaurussystemet, for produksjon av nye datasett/versjoner. • identifisere aktuelle registre hvor NGT bør registreres (ved metadata). DIFIs *data.norge.no* er selvskreven, men det kan også være andre, eksempelvis Termportalen[1] og Språkrådets oversikt over termlister og termbaser[2]
9.3	**Utarbeide dokumentasjon** Denne aktiviteten innebærer å utarbeide en helhetlig dokumentasjon som beskriver hvordan NGT skal brukes og vedlikeholdes, samt dens tekniske infrastruktur. Prinsipper/regler og avgjørelser som legges til grunn for vedlikehold og bruk av NGT må riktignok dokumenteres fortløpende gjennom hele prosjektet som del av alle relevante aktiviteter. Det vi her snakker om er å omdanne slik fortløpende dokumentasjon til en helhetlig og konsistent «pakke» som skal fungere også for framtidige brukere, redaktører og systemadministratorer.
9.3.1	Utarbeide NGT håndbok **Ansvarlig:** Prosjekt- og koordineringsgruppe (hovedsakelig tesaurusfaglig koordinator)

[1] http://www.terminologi.no/forside.xhtml - en nasjonal portal for terminologi
[2] http://www.sprakradet.no/Tema/Terminologi-og-fagspraak/Lenker/

Aktivitet	*Beskrivelse*
	Utarbeide håndbok for vedlikehold og oppdatering av NGT, inkludert • retningslinjer for valg, strukturering og koding av begreper • beskrivelse av arbeidsflyt for oppdatering av NGT
9.3.2	Utarbeide indekseringsregler **Ansvarlig:** Prosjekt- og koordineringsgruppe (hovedsakelig tesaurusfaglig koordinator) Utarbeide retningslinjer for bruk av NGT til indeksering, i første omgang med tanke på manuell indeksering.
9.3.3	Utarbeide systemdokumentasjon **Ansvarlig:** Prosjekt- og koordineringsgruppe (hovedsakelig utvikler/teknisk koordinator) Utarbeide dokumentasjon av infrastrukturen NGT vedlikeholdes i. Beskrivelsen bør omfatte • Datamodell for NGT • Metadataformat for datasettene som publiseres • Beskrivelse av den tekniske løsningen NGT vedlikeholdes i (arkitektur) • Fasiliteter som angår interoperabilitet: Beskrivelser av eventuelle APIer, endepunkt, formater som NGT-data publiseres i, etc.
9.4	**Etablere driftsorganisasjon** **Ansvarlig:** Styringsgruppen Før NGT lanseres bør driftsorganisasjonen være på plass, se beskrivelse av denne i kapittel 11. I dette inngår også kompetanseoverføring fra prosjektet til driftsorganisasjon.
10	**Lansering av NGT**
	Ansvarlig: Prosjekt- og koordineringsgruppen Aktiviteten innebærer å: • Sørge for at tesaurussystemet er satt opp til å støtte arbeidsflyten for oppdatering av NGT (fra forslag (om endringer eller nye begrep) til godkjenning) • Publisere NGT 1.0, også som nedlastbart datasett, se aktivitetene 9.1 og 9.2. Registrere datasettene i vedtatte registre • Promotere og arrangere selve lanseringen, gjerne i form av et seminar.
11	**Kommunikasjon utad / informasjonsarbeid**

Aktivitet	*Beskrivelse*
	Ansvarlig: Prosjekt- og koordineringsgruppen Det er ønskelig å kjøre prosjektet som en mest mulig åpen prosess, men løpende og god kommunikasjon med brukermiljøer og andre interessenter. Plan for dette bør lages tidlig i prosjektet. Ulike kanaler og fora kan vurderes: • Nettsted • Blogg • Sosiale medier (Facebook-gruppe) • Seminarer, eksempelvis ved lansering • Delta med publikasjon/foredrag på konferanser, i Norge og internasjonalt • Artikler i populærtidsskrifter, eksempelvis Bok og Bibliotek og SLQ Det er viktig å legge kommunikasjonen på et realistisk nivå som kan opprettholdes gjennom hele prosjektet.
12	**Prosjektledelse** **Ansvarlig:** Prosjektleder Prosjektledelse innebærer å ha det overordnete ansvar for at prosjektets resultatmål nås og at det gjennomføres etter vedtatte planer. I dette inngår å planlegge, koordinere og følge opp aktiviteten i prosjektet, dvs. sørge for systemer som oppdager avvik, løse problemer som oppstår underveis. Det skal rapporteres til Styringsgruppe med fastsatte intervaller og ved avvik.
12.1	**Prosjektrapport** **Ansvarlig:** Prosjektleder Det skal skrives en rapport som oppsummerer prosjektets arbeid og formidler erfaringer samlet gjennom prosjektet. Spesielt er det viktig å beskrive valgt metode for integrering av vokabularene, erfaringer med denne og tanker om mulige forbedringer/tilpasninger. Dette fordi det kan være aktuelt å integrere flere vokabularer inn i NGT på et senere tidspunkt.



Figur 2 Arbeidsflyt for utvikling av NGT 1.0.

7.1 Tidsplan for utvikling av NGT 1.0

Aktivitetene beskrevet ovenfor er her tidsplanlagt i en abstrakt kalender. Slik prosjektet er estimert vil prosjektperioden vare i 2.5 år.

Arbeidet er vanskelig å beregne omfanget av, blant annet fordi det er usikkert hvor god systemstøtte vi vil oppnå for selve integreringsarbeidet (Aktivitet 7). Samtidig er det denne aktiviteten som representerer den største delen av arbeidet. Det er derfor tilstrebet et konservativt estimat, basert på at integreringen av emneordene i hovedsak må gjøres manuelt (med redigeringsstøtte). Ressursbehov er beregnet ut fra et estimert ressursbehov per term, som forklart nedenfor.

7.1.1 Arbeidsmengde og tidsbruk i Aktivitet 7 Integrering

Et vokabular V som skal integreres i NGT har som regel et visst overlapp med NGT, slik denne er når V skal innlemmes. Dette overlappet er estimert konservativt. Termene i overlappet må sjekkes for å avdekke homonymi o.a., men dette er mindre krevende enn å «innplassere» termer som ikke finnes i NGT fra før.

La *T* være *antall minutter som brukes for å «innplassere» en term fra et vokabular inn i NGT*. I dette arbeidet inngår også å rådføre seg med fagspesialister der dette er nødvendig. *T* er estimert ut fra forsøket med NGT 0.1 (se Appendiks 3) og fra integreringsarbeidet beskrevet i (Nilbe 2012).

I våre beregninger brukes

- *T*=4 for termer som ikke finnes i NGT på integreringstidspunktet
- *T*=2 for termer som finnes i NGT fra før (overlapp). I slike tilfeller skal det bare sjekkes at det virkelig dreier seg om samme begrep

Hvert vokabular *V* er tilordnet en kompleksitetsfaktor K_v som uttrykker antatt vanskelighetsgrad av *V*. Denne brukes bare for ikke overlappende termer.

Tidsbruk pr term for vokabular *V* blir da $T * K_{v.}$

Gitt *antall termer i hvert vokabular* og estimat av *overlappet med NGT på integreringstidspunktet*, kan vi beregne den totale arbeidsmengden (f.eks. i timeverk) knyttet til innlemming av dette vokabularet.

Ut fra *antall personer* (FTE[1]) som skal integrere *V*, finner vi *varigheten* på arbeidet. Før utlegg i tidsplanen er varigheten forlenget med ca. 50%. Dette fordi vi antar at ingen kan jobbe intensivt med denne type arbeid hele dagen hver dag.

Eksempel:

> Realfagstermer er tilordnet en kompleksitetsfaktor 2.0 for innplassering av termer i NGT.

[1] Full time equivalents

Da blir estimert tidsbruk pr term: *4 * 2 = 8 minutter*

Vi regner at Realfagstermer inneholder ca 12750 termer som ikke overlapper med NGT på integreringstidspunkt.

Dette gir *12750*8/60 = 1700 timeverk*.

Med 2 personer på jobben blir varigheten 850 timer eller *5,3 mnd*.

Tilsvarende beregning for overlappende termer gir *0.2 mnd*.

Dette blir totalt *5.5 mnd*, som i tidsplanen er strukket ut til 9 mnd.

På samme måte er integreringsarbeidet beregnet for alle vokabularene. Antall personer på hvert vokabular er satt til 2 for alle bortsett fra NBs emneordslister, hvor det på grunn av volumet er forutsatt 4 personer i tidsberegningen.

				År 1												År 2												År 3											
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	Etablere prosjektet																																				
		Formalisere prosjektet																																			
		Danne prosjektorganisasjon																																			
2	Etablere teknisk infrastruktur og systemstøtte																																				
		Anskaffe og implementere tesaurussystem																																			
		Kvalitetssikre representasjonsmodell																																			
		Anskaffe eller utvikle støtte for integrering av vokabularer																																			
		Anskaffe publiseringsplattform for data																																			
3	Preprosessering av de enkelte vokabularer																																				
		Preprosessere Humord - etablere NGTs starttesaurus NGT0																																			
		Preprosessere juridiske og menneskerettslige emneord																																			
		Preprosessere Realfagstermer																																			
		Preprosessere Nasjonalbibliotekets emneordslister																																			
4	Avklar håndtering av begreper innenfor helsefag og psykologi i eksisterende vokabularer																																				
		Utred håndtering av begreper innenfor helsefag og psykologi i eksisterende vokabularer																																			
		Bestem hvordan begreper innenfor helsefag og psykologi skal håndteres i NGT																																			
5	Utfør vedtak om begreper innenfor helsefag og psykologi																																				
6	Utarbeide toppstruktur for NGT																																				
7	Integrere de preprosesserte vokabularene med NGT																																				
		Integrere Juridiske og menneskerettslige emneord inn i NGT0																																			
		Integrere Realfagstermer inn i NGT																																			
		Integrere NBs emneordslister inn i NGT																																			
8	Helhetlig gjennomgang og kvalitetssikring av NGT																																				
9	Klargjøring av NGT for «markedet»																																				
		Avklar juridiske forhold omkring opphavsrett og bruksrett																																			
		Klargjør for publisering av datasett																																			
		Utarbeid dokumentasjon																																			
			Utarbeid NGT håndbok																																			
			Utarbeid indekseringsregler																																			
			Utarbeid systemdokumentasjon																																			
		Etablér driftsorganisasjon																																			
10	Lansering av NGT																																				
11	Kommunikasjon utad / informasjonsarbeid																																				
12	Prosjektledelse																																				
		Skrive prosjektrapport																																			

8 Utvikling av NGT 1.0 – Prosjektorganisasjon

I det følgende beskrives de ulike enhetene og deres oppgaver i foreslått prosjektorganisasjon for utviklingen av NGT 1.0, se Figur 3 for visualisering. Kompetansebehov for og bemanning av enhetene er beskrevet i 8.4.

8.1 Prosjekt- og koordineringsgruppe.

Prosjektgruppen koordinerer alt arbeid, identifiserer oppgaver og problemstillinger som er felles for alle arbeidsgruppene, og søker felles løsninger for disse. Prosjektleder rapporterer til Styringsgruppen regelmessig og ved behov. Prosjektgruppen bør bestå av:

- Prosjektleder
- Utvikler(e)/teknisk koordinator
- Ansvarlige for de til enhver tid aktive arbeidsgruppene omtalt i 8.2 og 8.2.3 nedenfor
- Tesauruskoordinator

Prosjektgruppen har også ansvar for den totale kvaliteten til NGT som kunnskapsorganisatorisk ressurs. Dette innebærer å

- ha overblikket over helheten i NGT, m.a.o. se på tvers av hierarkiene og fjerne inkonsistens og redundans som kan oppstå når ulike grupper jobber med hver sine deler av NGT
- ivareta konsistens når det gjelder formalia, eksempelvis rettskriving, valg mellom ulike ordformer (f.eks. entall/flertall), og annet
- utarbeide tesaurusfaglig dokumentasjon (hovedsakelig NGT håndbok og indekseringsregler) og tilrettelegge for publisering av NGT som datasett
- sørge for hensiktsmessig infrastruktur og systemstøtte til arbeidsgruppene

Totalansvaret bør fordeles mellom prosjektleder, utvikler og tesauruskoordinator etter hva som er naturlig ut fra den enkeltes rolle og kompetanse.

8.2 Oppgavespesifikke, fleksible arbeidsgrupper

De ulike aktivitetene har ulike behov for kompetanse, og det vil bli behov for å danne midlertidige arbeidsgrupper med spesifikke kompetanseprofiler for noen av disse. For hver arbeidsgruppe utnevnes en ansvarlig person som i arbeidsgruppens virkeperiode deltar i prosjekt- og koordineringsgruppen. Dette for å sikre at felles problemstillinger oppdages og løses på tvers av arbeidsgruppene.

8.2.1 Arbeidsgrupper knyttet til vokabularene som skal integreres

Disse arbeidsgruppene skal jobbe med preprosessering av hvert enkelt vokabular, før selve integreringen i NGT. Fire grupper etableres for dette:

- Arbeidsgruppe for prosessering av Humord
- Arbeidsgruppe for prosessering av juridiske og menneskerettslige emneord

- Arbeidsgruppe for prosessering av Realfagstermer
- Arbeidsgruppe for prosessering av emneord fra NB

8.2.2 Arbeidsgrupper knyttet til de ulike hierarkiene i NGT

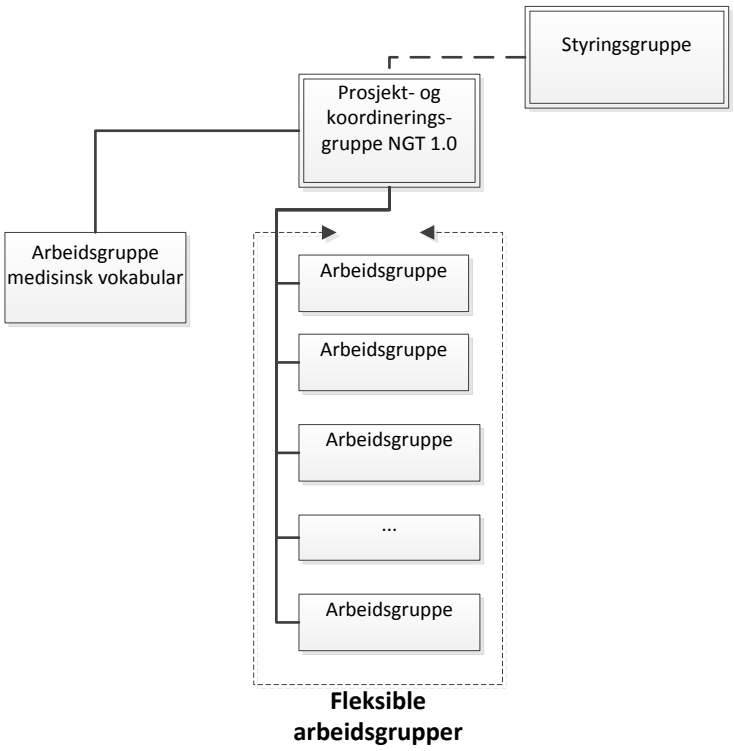
Disse arbeidsgruppene skal jobbe med integreringen av hvert enkelt vokabular inn i NGT. Gruppene vil være inndelt etter emne og jobbe med hver sine deler av NGT. I hovedsak vil disse bli bemannet av de samme personene som bemanner arbeidsgruppene nevnt i 8.2.1.

8.2.3 Arbeidsgruppe medisinsk vokabular

Denne arbeidsgruppen skal utrede og anbefale hvordan de medisinske begrepene skal håndteres i NGT i lys av norsk oversettelse av MeSH. Se kapittel 7, Aktivitet 4.

8.3 Styringsgruppe

Prosjektet bør ha en styringsgruppe som følger utviklingen og tar beslutninger om eventuelle kursendringer. Prosjektleder rapporterer til styringsgruppen med fastsatte mellomrom, og ved avvik.



Figur 3 Prosjektorganisasjonen

8.4 Bemanning av prosjektorganisasjonen

8.4.1 Bemanning av Prosjekt- og koordineringsgruppen

De mest sentrale personene må være dedikert til prosjektet, dvs. det vil neppe fungere å ha personer som er bundet opp i mange andre oppgaver. I tillegg er det viktig at både NB og UBO er representert blant de faste medlemmene i gruppen.

Vi foreslår derfor at

- Prosjektleder ansettes på full tid ved NB på prosjekt for en periode på 3 år med arbeidssted Oslo
- Utvikler/teknisk koordinator leies inn etter anbud eller ansettes midlertidig ved NB på prosjektet for en periode på 3 år med arbeidssted Oslo
- Tesauruskoordinator er knyttet til UBO. Også denne bør være på tilnærmet full tid, i hvert fall ha NGT som hovedbeskjeftigelse i utviklingsperioden

Det framgår av tidsplanen at prosjektet er beregnet til 2,5 år. Når vi likevel foreslår å ansette kjernepersonell for 3 år, er det med tanke på arbeid etter lansering. Et sentralt punkt her er overføring av kompetanse og ansvar til driftsorganisasjonen (hovedsakelig redaksjonsgruppen). Det tas også høyde for usikker beregning av ressursbehov.

Kompetansebehov

Prosjektleder må ha dokumentert bakgrunn/erfaring i prosjektledelse, og ha en solid faglig bakgrunn i kunnskapsorganisering.

Utvikler/teknisk koordinator må ha god kunnskap om og erfaring i XML, utvikling for web, semantisk web og linked data, helst anvendt på bibliotekfeltet. Vedkommende bør ha god oversikt over standarder, metoder og verktøy på feltet. Vedkommende bør ha erfaring med metadata, og kjennskap til språkteknologi er en fordel.

Tesauruskoordinator må ha god kunnskap om emnesystemer generelt og tesauruser spesielt, praktisk så vel som teoretisk. Vedkommende bør også ha gode kunnskaper om linked data og semantisk web. En viktig egenskap er også evne til å tenke helhet og felles beste, samtidig som faglig stolthet og historie knyttet til hvert enkelt vokabular respekteres.

8.4.2 Arbeidsgruppe for vokabular innen helsefag og psykologi

Denne gruppen bør totalt bemannes av 4-5 personer og ledes av en person med kompetanse på emneordsbruk/kunnskapsorganisering innenfor helsefag – både når det gjelder standarder, emnesystemer og praksis. En slik person kan sannsynligvis finnes på UBO Medisinsk bibliotek. De øvrige gruppemedlemmene må hentes blant de samme personene som bemanner arbeidsgruppene nevnt i 8.2 men valgt ut fra kunnskap om og interesse for helsefag/psykologi og befatning med emneord på området. Konkret betyr dette at både Realfagstermer, NORART og Humord bør være representert i gruppen.

8.4.3 Oppgavebaserte, fleksible arbeidsgrupper

Vi ser for oss at alle gruppene bemannes av en relativt begrenset og stabil mengde (totalt 10-15) bibliotekansatte fra UBO, NB eller andre institusjoner i dagens Humord-samarbeid, som

- til sammen representerer deltakerinstitusjonene (NB og UBO), NGTs faglige nedslagsfelt og de opprinnelige vokabularene mest mulig jevnt
- kan inndeles i grupper på ulik måte etter behov

Kompetansemessing må alle her ha solid faglig bakgrunn i kunnskapsorganisasjon. Det er ønskelig at noen også har gode kunnskaper om IT/semantisk web/ åpne lenkede data.

For preprosesseringen av vokabularene (se 8.2.1og kapittel 7, Aktivitet 3) er det av avgjørende betydning at gruppemedlemmene kjenner vokabularet inngående, - det er derfor naturlig å bemanne disse gruppene hovedsakelig med personer som har jobbet praktisk med vokabularet. Hvis mulig, bør det også delta personer som står litt fjernere fra det enkelte vokabular.

Vi regner med at disse også innehar tilstrekkelig faglig bakgrunn innenfor sine felt til å bemanne gruppene som skal arbeide med de ulike hierarkiene i NGT (se kapittel 7, Aktivitet 7), og som skal ha en emnebasert profil.

Kompetansebehov

Kjernen i alle arbeidsgruppene skal være personer med bibliotekfaglig/kunnskaps-organisatorisk kompetanse og kunnskap om et eller flere av vokabularene. Det er også ønskelig at eventuelle utviklere som har jobbet konkret med det enkelte vokabular kan delta i gruppene etter behov, i samarbeid med prosjektets utvikler/teknisk koordinator.

I tillegg må det være lett tilgang til spesialkompetanse i form av fagreferenter/ forskningsbibliotekarer innenfor de ulike emneområdene som NGT dekker. Det er ønskelig at disse er spesielt interessert i terminologi og begrepsdannelse innenfor sitt fagområde.

8.4.4 Styringsgruppen

Kjernen i Styringsgruppen bør være representanter fra kunnskapsorganisasjonsfaglig ledelse ved UBO og NB, henholdsvis. Leder for gruppen bør komme fra NB. I tillegg bør andre interessenter – i første omgang deltakerinstitusjonene i Humord-samarbeidet – inviteres til å delta som observatører. Dette for å oppnå forankring i en noe videre sirkel enn UBO og NB.

8.5 Eksterne ressurser

I tillegg til selve prosjektorganisasjonen, som omfatter de som jobber fast i prosjektet over tid, vil det bli nødvendig å innhente kompetanse utenfra på enkelte områder.

8.5.1 Fagspesialister

Som nevnt ovenfor er det avgjørende å ha tilgang til fagkunnskap innenfor de ulike områder som NGT dekker. I praksis betyr dette at fagreferenter (UBO) og forskningsbibliotekarer (NB) må kunne involveres ved behov, både for avklaring av konkrete problemstillinger og for gjennomgang av enkelte deler av tesaurusen.

Behovet for dette vil variere mellom de ulike deler av NGT, og uttak av slike ressurser, som må initieres og styres av arbeidsgruppene, vil kunne bli ganske ad hoc. Vi har derfor valgt ikke å inkludere fagspesialistene i selve prosjektorganisasjonen.

8.5.2 Juridisk kompetanse

Før NGT kan publiseres må det avklares under hvilke betingelser dette skal skje, se kapittel 7, Aktivitet 9.1. Her vil prosjektet måtte konsultere jurist(er) med kunnskap om opphavsrett og åpne data.

8.5.3 Annet

Det vil kunne oppstå behov for å trekke inn eksterne fagmiljøer i noen diskusjoner, uten at vi foreløpig kan konkretisere dette i særlig grad. For utredningen om den helsefaglige- og psykologiske delen av NGT er imidlertid naturlig å rådføre seg med MeSH-miljøet ved Helsebiblioteket.no.

9 Utvikling av NGT 1.0 – Verktøy og ressurser

I dette kapitlet beskrives ulike ressurser som trengs for å utvikle NGT 1.0, nærmere bestemt infrastruktur og ulike kunnskapsressurser.

9.1 Infrastruktur

9.1.1 Programvare for tesaurusforvaltning

For å oppnå en effektiv forvaltning av NGT er det viktig å ha et godt system som støtter hele arbeidsprosessen for drift av NGT.

Det finnes en del systemer som retter seg mot forvaltning av tesauri og lignende vokabularer, men ikke svært mange. I stedet for å spesifisere en liste med detaljerte krav til et slikt system fra begynnelsen av, valgte vi derfor å ta utgangspunkt i noen eksisterende systemer som alle virker aktuelle, og sammenligne disse med hensyn på i underkant av 50 parametre, inndelt i følgende grupper:

- Hvordan systemet støtter aktivitetene som inngår i å drifte en tesaurus: Modellering av tesaurusen, innsamling og strukturering av begreper fra ulike kilder, oppdatering av begrepene samt deling av tesaurusen via eksport, publisering av datasett og/eller tilgjengelig endepunkt (f.eks. SPARQL)
- Brukergrensesnitt
- Administrasjon av brukere
- Tilleggsfunksjonalitet
- Ikke-funksjonelle egenskaper

Systemene er valgt ut etter egen kunnskap og ved å konsultere nettressurser med oversikt over slike systemer, for eksempel en oversikt på nettsiden til American Society for Indexing[1].

Til sammen seks systemer ble vurdert:

- MultiTes[2] fra Multisystems, Florida, USA
- Poolparty[3] fra Semantic Web Company (SWC), Wien
- Synaptica[4] fra Synaptica LLC, Colorado, USA
- TemaTres[5] fra R020 Bibliotecología y ciencias de la informacíon, Argentina
- Thesaurus Master®[6] fra DataHarmony, Division of Access Innovations, Inc., New Mexico, USA
- VocBench[7], utviklet av Food and Agriculture Organization of the United Nations (FAO) og ART Group (Artificial Intelligence Research at Tor Vergata) of the University of Rome 'Tor Vergata'

Nedenfor presenteres vurderingen av hvert system kort. En mer fullstendig beskrivelse finnes i Appendiks 5.

- **MultiTes** har isolert sett det vi trenger for tesaurusredigering, men har ingen støtte for andre ting, som f.eks. sammenligning av ulike tesauri. Det har også et noe gammelmodig grensesnitt/utseende.
- **Synaptica KMS** ser ut til å være et fremtidsrettet system som har det vi trenger for utvikling av NGT. Positivt er også støtte for samarbeid i grupper. Ekstrafunksjonalitet som utvikles (f.eks. ved IMS) virker å være orientert mot virksomhetsinternt innhold, og vil trolig ikke være så nyttig for NGT. Det er uvisst hvordan tesaurusen kan presenteres for sluttbruker. Eksemplene på publisering med Synaptica Publication Suite viser bare alfabetisk oppslag. En svakhet er også at SPARQL ikke støttes. Likevel, Synaptica framstår som et fleksibelt og brukervennlig system for selve tesaurusforvaltningen.
- **TemaTres** inneholder mye bra funksjonalitet, og har isolert sett antakelig det meste av det vi trenger for NGT. Det er likevel en del negative trekk, for eksempel at datamodellen ikke skiller mellom begrep og term, måten å løse flerspråklige tesauri på og mangel på støtte for rollebasert tilgang. Mye av dokumentasjonen på nett er på spansk, og er for en stor del fra 2011. Programvaren i seg selv utvikles imidlertid jevnt, nyeste versjon (1.81) var publisert høst 2014. Det er vanskelig å finne informasjon om ansvarlig firma R020, men siden copyright til TemaTres holdes av en enkeltperson, tyder det

[1] http://www.asindexing.org/about-indexing/thesauri/thesaurus-management-software/
[2] http://multites.net/index.htm
[3] http://www.poolparty.biz/
[4] http://www.synaptica.com/
[5] http://www.vocabularyserver.com/
[6] http://www.dataharmony.com/services-view/thesaurus-master/
[7] http://vocbench.uniroma2.it/

på at firmaet er lite (og dermed sårbart). Alt i alt, som produkt betraktet virker hele opplegget litt umodent.

- **Thesaurus Master®** alene virker å være et standard tesaurussystem fra en solid leverandør som sannsynligvis har det meste av basisfunksjonaliteten som NGT trenger, men heller ikke mer. Det nevnes for eksempel ingen støtte for sammenligning/mapping mellom vokabularer. Det mest spennende med DataHarmony-verktøyene er for øvrig indekseringsverktøyet M.A.I.™, men dette blir foreløpig sekundært i NGT-sammenheng. Koblet med en høy pris blir det derfor ikke så relevant.
- **VocBench** har de nødvendige støttefunksjonene for å forvalte en tesaurus på en distribuert måte, og med en kontrollert arbeidsflyt. Erfaring viser at det er noen tekniske utfordringer med å bruke gratisversjonen av det anbefalte repositoriet (GraphDB Lite), men vi må regne med at dette bedrer seg etter hvert. Det er også fullt mulig å gjenbruke konfigurasjonen for EuroVoc[1] basert på Virtuoso. En negativ ting er at enkel SKOS[2] ikke støttes, (bare SKOS-XL), men dette ligger i planene, likeså fasiliteter for sammenligning av vokabularer.

Konklusjon

Slik vi ser det nå, er verken TemaTres eller Thesaurus Master® aktuelle, - førstnevnte på grunn av manglende modenhet og et sårbart utviklingsmiljø, sistnevnte på grunn av høy pris i forhold til funksjonalitet.

Av de andre er både PoolParty, Synaptica og MultiTes aktuelle. Av disse er MultiTes prismessig den rimeligste, men absolutt minst avansert, da er det ikke er mulighet for støtte til sammenligning/lenking mellom vokabularer gjennom verktøyet. PoolParty og Synaptica er begge avanserte, tilbyr mange fasiliteter for lenkede data og kan gi oss mye, men til en høy pris.

VocBench er derimot fri programvare, som vi nå til en viss grad har prøvd ut i NGT 0.1 (se Appendiks 3) og som vi ser har potensiale til å bli bra. Her har vi også mulighet til å påvirke utviklingen og om ønskelig utvikle egne tilleggsmoduler. I forhold til et rent kommersielt verktøy med betalt vedlikehold, er det imidlertid mer krevende å ta i bruk fri programvare som VocBench, da installering (av VocBench så vel som nødvendige tilleggsprogramvare), konfigurering og generelt oppsett må gjøres av den enkelte brukerinstitusjon. Bruk av VocBench forutsetter derfor at vi har teknisk kyndige personer i prosjektet under hele utviklingen, selv om det er god støtte å finne gjennom brukergruppen vocbench-user og konkret gjennom vår kontakt med EUROVOC-miljøet.

Alt tatt i betraktning, anbefaler vi å ta i bruk VocBench sammen med en triple-store som passer inn i NBs øvrige driftsmiljø.

[1] http://eurovoc.europa.eu/drupal/
[2] http://www.w3.org/2004/02/skos/

9.1.2 Representasjonsmodell

Det er naturlig å ta utgangspunkt i eksisterende standard for flerspråklige tesauri, ISO 25964 (International Standardization Organization 2011-2013). Denne definerer en datamodell og et XML-skjema som i prinsippet kan brukes til representasjon og utveksling av data. Støtten for denne standarden er imidlertid oftest ufullstendig, især hva angår format for import/eksport av data. Derimot er det lettere å finne støtte for den enklere modellen SKOS (Simple Knowledge Organization System) som er basert på RDF og utgjør en av W3C-standardene for semantisk web.

Både på grunn av den generelle utbredelsen, og på grunn av støtten hos verktøyet (VocBench) som vi har benyttet i sammenheng med NGT 0.1, har vi valgt å ta utgangspunkt i SKOS, med utvidelsen SKOS-XL. Sistnevnte gjør det mulig å knytte egenskaper (dato m.m.) til selve termene (skos-xl:Label).

SKOS/SKOS-XL kan bare delvis representere datamodellen i ISO 25964. Forholdet mellom SKOS og tesaurus-standarden fremgår av dokumentet *Correspondence between ISO 25964 and SKOS/SKOS-XL models*[1]. Dette inneholder også forslag til utvidelser av SKOS/SKOS-XL som skal favne semantikken i ISO 25964.

Vårt utgangspunkt har vært å representere HUMORD ved hjelp av SKOS/SKOS-XL med minst mulig tap av informasjon. De øvrige emneord-lister som inngår i NGT 0.1 er enklere og representerer ingen utfordringer med hensyn til representasjon. HUMORD blir jevnlig publisert i et (proprietært) XML-format. Detaljer og problemer knyttet til konvertering mellom dette og SKOS/SKOS-XL er nærmere beskrevet i Appendiks 4.

9.2 Kilder for begreper og termer

En god tesaurus skal til enhver tid representere oppdatert fagterminologi innenfor de emner den dekker. For å holde kvaliteten oppe gjennom hele livsløpet er gode kilder og hjelpemidler vesentlig.

ISO 25964 (International Standardization Organization 2011-2013.) skisserer følgende typer kilder til hjelp under utvikling og vedlikehold av tesauri (punktene er fra kapittel 13.1.3.3 Vocabulary resources, s. 89-90):

a) Eksisterende tesauri eller klassifikasjonsskjema innenfor samme fagområder. Disse kildene kan gi ideer til termer, struktur eller begge deler
b) Samlinger av termer eller "ofte stilte spørsmål", gjerne slike som kolleger har samlet inn gjennom sitt arbeid
c) Register til databaser eller andre relevante referanseverktøy
d) Søkelogger, for eksempel logger som viser hvilke søketermer brukerne oftest bruker

[1] http://www.niso.org/apps/group_public/download.php/12351/Correspondence%20ISO25964-SKOSXL-MADS-2013-12-11.pdf

e) Standard referanseverker som ordbøker, encyklopedier, terminologier. Disse er nyttige først og fremst for å definere og avgrense termers betydningsomfang, og ikke nødvendigvis for valg av termer.

De ulike kildene kan brukes på ulike måter. Noen sentrale funksjoner er:

- Gi forslag til nye begrep som bør inkluderes
- Som hjelp til å definere termenes begrepsmessige innhold
- Hjelp til valg av term som skal være hovedemneord og eventuelle synonymer
- Hjelp til korrekt språklig form og rettskriving
- Hjelp til faglig struktur og samstemming av tesaurusen mot andre, tilsvarende fagressurser

Typer av kilder og deres roller i arbeidet med NGT:

- Andre vokabularer. Dette kan være kunnskapsorganisatoriske ressurser som tesauri, klassifikasjonsskjema, emneregistre, taksonomier og emneordslister. Eksempler: AGROVOC og MeSH. Det kan også omfatte terminologiske ressurser som f.eks. termbanker. Eksempel: Snorre og Norsk ordvev. Roller: forslag til termer og faglig struktur
- Ordbøker innen alle språkformene tesaurusen skal dekke. Eksempel: Bokmålsordboka. Roller: forslag til termer, rettskriving og definisjoner
- Leksika, både generelle og fagspesifikke. Eksempel: Store norske leksikon, Store medisinske leksikon. Roller: forslag til termer, definisjoner og faglig struktur
- Fulltekstressurser som institusjonsarkiver og e-tidsskrifter. Eksempler: DUO, DOAJ og Idunn. Nyttig kilde til fagspråk (fritt skrevet fagstoff). Roller: forslag til termer
- Kvalitetssikrete nettsteder som er strukturert og formulert av fagmiljøer. Det kan være kvalitetssikrete nettressurser, temasider, ofte stilte spørsmål. Eksempel: Språkrådets nettsider. Roller: forslag til termer og faglig struktur
- Søkelogger. Disse kan gi tilgang til brukerformulerte spørsmål og søkeord. Det kan være en nyttig kilde for ordtilfang, men også for å kartlegge emner som ikke er uttrykt eller som er mangelfullt uttrykt. Eksempel: søkelogg fra BIBSYS. Roller: forslag til termer
- Fagpersoner. Eksempler: Fagreferenter på universitetsbibliotekene og vitenskapelig personale ved universitetene. Disse vil være nødvendige kilder for kvalitetssikring av begreper, termer og faglig struktur. Roller: forslag til termer, faglig struktur og definisjoner.
- Eksisterende retningslinjer i ISO-standarder, fagbøker mm. Humord håndbok er her en sentral ressurs. Roller: Kilde for beste praksis når det gjelder tesaurusens språklige form og struktur.

10 Utvikling av NGT 1.0 – Finansiering

Arbeidet som fører fram til NGT 1.0 er relativt ressurskrevende. Det er ikke estimert noen total kostnad for prosjektet, men slik prosjektplanen nå foreligger vil følgende representere direkte kostnader:

- Ansettelse av 2 personer for 3 år
- Anskaffelse av nødvendig infrastruktur. Foreslått verktøy er fri programvare, men det må tas høyde for innkjøp at kommersielle tilleggsprodukter og annet, f.eks. database

Selve utviklingsarbeidet – spesielt integrering av de opprinnelige vokabularene – krever svært spesiell kompetanse, og vi ser ikke for oss at det kan leies inn personer til dette. Tvert imot vil arbeidet måtte utføres av operativt personell ved NB og UBO, hvorav flertallet vil komme fra UBO. Alle som skal delta i arbeidet med vokabularene må kunne arbeide konsentrert med dette over lengre perioder, og må derfor fritas fra andre oppgaver.

Følgende finansiering foreslås:

- NB, som framtidig eier og forvalter av NGT bør bekoste selve prosjektlederfunksjonen, dvs. ansettelse av de to personene for 3 år, anskaffelse og vedlikehold av infrastruktur, bl.a. programvare.
- NB skal også bekoste innsatsen til øvrig personale i NB.
- NGT 1.0 er tenkt å være en nasjonal ressurs. Vi mener derfor det er naturlig å kanalisere utviklingsmidler[1] inn i prosjektet, som kan finansiere UBOs innsats på vanlige prosjektvilkår. Dette betyr at UBO forventes å yte en viss egeninnsats, og stille eget fagpersonale til rådighet i rimelig grad. Midlene bør bevilges utenfor den ordinære søknadsrunden, slik at finansieringen blir forutsigbar for hele perioden.

11 Drift av Norsk generell tesaurus – etter versjon 1.0

11.1 Aktører

Som det framgår av kapittel 8 vil tesaurussamarbeidet fram til NGT 1.0 er etablert bestå av NB og UBO. Humord-samarbeidspartnerne må få anledning til å uttale seg i spørsmål knyttet til etablering og organisering, men UBO kan sannsynligvis representere disse institusjonene og ivareta deres interesser i denne fasen.

Den første driftsorganisasjonen etter lansering av NGT 1.0 bør være sentrert i NB med UBO og minst en institusjon fra Humord-samarbeidet som partnere.

Senere, i tråd med den videre utvikling av NGT, kan det tenkes at nye partnere kommer med. Eksempelvis, ved integrering av nye vokabularer må det vurderes om

[1] Her menes utviklingsmidler forvaltet av NB (http://www.nb.no/Bibliotekutvikling)

eierinstitusjonen skal delta i det videre arbeidet med NGT på sitt felt. Dette vil avhenge av vokabularets størrelse og betydning, samt eierens interesse. Det er viktig at antall medlemmer i redaksjonsgruppen (se nedenfor) er tilpasset termtilfanget, er fleksibelt og kan variere over tid etter behov.

11.2 Roller, myndighet og ansvar

Etter at en formell, skriftlig avtale om samarbeid om NGT er inngått mellom NB og UBO ved etablering av versjon 1.0, må arbeidet deretter organiseres på permanent basis.

Det er, som understreket på møtet med andre bibliotekinstitusjoner 30. april 2014, NB som bør være eier av og hovedansvarlig for Norsk generell tesaurus. Slik er også situasjonen i andre land. Som en vesentlig bidragsyter og utvikler av den tesaurusen samarbeidet skal tuftes på, bør UBO (også på vegne av Humord-samarbeidspartnere) ha en sterk faglig og administrativ tilknytning til tesaurusen også i fortsettelsen.

Tesaurusen må styres i tråd med følgende prinsipper:

- Arbeidet må være minst mulig byråkratisk organisert
- Emneordsarbeidet må ha en jevn framdrift, med kortest mulig utredningstid
- Fagreferentene[1] må trekkes sterkt inn i arbeidet

Vi ser for oss at fagreferenter som hovedregel er ansvarlig for det faglige innholdet i hierarkiene. I enkelte hierarkier som er generelle eller som er innenfor bibliotek- og informasjonsvitenskap, kan bibliotekarer være ansvarlig for det faglige innholdet.

11.2.1 Faggrupper - med ansvar for hver sine hierarkier

Ansvaret for de ulike hierarkiene som er utviklet/skal utvikles, spres ut på enheter i UBO, hos samarbeidspartnerne i Humord-samarbeidet, eventuelle nye samarbeidspartnere samt NB i tråd med kompetanse/spesialiseringsgrad i den enkelte institusjon. I tråd med retningslinjene i forrige avsnitt etableres det grupper av fagreferenter (om ønskelig supplert med bibliotekarer) eller bibliotekarer for hvert hierarki. Det kan være aktuelt at én gruppe er ansvarlig for flere hierarkier. Antall personer i gruppene kan variere. Den enheten som har fått det faglige ansvaret, organiserer arbeidet i tråd med prinsippene i kulepunktene over. Om det skal trekkes inn fagfolk fra flere institusjoner vurderes av gruppene.

Faggruppene har følgende oppgaver:

- Motta og behandle forslag til termer fra brukermiljøene (basert på dokumenttilfang i disse)
- Foreslå nye termer i tråd med egen tilvekst
- Gjennomgå emneordsvokabularer som det er aktuelt å innlemme i tesaurusen
- Foreslå rettelser og tilføyelser i eksisterende hierarkier

[1] Betegnelsen *fagreferent* omfatter her også personer med tilsvarende rolle i NB, selv om stillingsbetegnelsen der er en annen (*forskningsbibliotekar*).

- Foreslå synonymer, nærsynonymer, hierarkiske eller assosiative relasjoner mellom termer
- Holde fagspråket oppdatert
- Følge de til enhver tid gjeldende indekseringsregler
- Delta i debatter/møter i regi av NB som angår tesaurusen
- Samarbeide tett med redaksjonsgruppen.

Denne måten å organisere arbeidet på sikrer at arbeidet knyttes til operative miljøer med spesialkunnskap på feltet. Dette vil forhåpentligvis styrke det faglige eierskapet til tesaurusen og muligens også påvirke bruken av tesaurusen positivt.

11.2.2 Redaksjonsgruppe ledet av NB

Det er ikke tenkt at faggruppene selv skal innlemme termer i tesaurusen. Dette arbeidet utføres av en redaksjonsgruppe. Redaksjonsgruppen ledes av NB, men UBO skal være sterkt representert.

Det vil sannsynligvis være behov for møter mellom redaksjonsgruppen og faggruppene årlig eller oftere (særlig i starten) for å sikre at arbeidet blir konsistent, samarbeidet fungerer godt og at arbeidet følger oppsatte framdriftsplaner.

Redaksjonsgruppens ansvar er:

- Koordinere løpende vedlikeholdsarbeid
    - Sørge for at hierarkiene utvikles og holdes ved like
    - Sørge for at arbeidet med tesaurusen følger fastsatt framdriftsplan
    - Koordinere arbeidet i faggruppene og mellom de ulike redaktørene i redaksjonsgruppa
        - Fordele ansvaret for de enkelte hierarkiene
    - Publisere nye termer fortløpende
    - Sørge for at tesaurusen publiseres som åpne data
- Holde kontakt med brukermiljøene
    - Arrangere møter med brukere, potensielle brukere og andre indekseringsmiljøer
    - Markedsføre tesaurusen
    - Inngå avtaler om innlemmelse/samarbeid/mapping til andre, eksisterende vokabularer
- Være kompetansemiljø for tesaurusutvikling
    - Holde seg faglig oppdatert på "beste praksis" innen feltet
- Identifisere behov for større innsatser, som må organiseres utenfor det vanlige vedlikeholdet
    - Sette i gang spesielle utredninger

11.2.3 Styringsgruppe

Redaksjonsgruppa bør rapportere til en styringsgruppe bestående av representanter fra NB og UBO (evt. også med Humord-samarbeidspartnere), andre deltakere i

samarbeidet, representanter fra brukermiljøene, muligens også fra andre indekseringsmiljøer og andre aktuelle parter. Styringsgruppa møtes 1-2 ganger i året avhengig av behov.

11.3 Finansiering/kostnadsfordeling

Så snart NGT 1.0 er kommet i drift, bør NB dekke alle kostnader knyttet til sekretariatsfunksjon, utviklingsverktøy, reisevirksomhet i forbindelse med møter mellom faggrupper og andre kostander som følger naturlig av at NB er primus motor i samarbeidet.

De andre deltakerinstitusjonene i driftsorganisasjonen bekoster selv egen arbeidsinnsats i prosjektet, både på fagreferentsiden, av bibliotekfaglig personale og deltakere i redaksjonsgruppen.

Dersom/når ytterligere vokabularer skal innlemmes, er det naturlig at eierne av disse, bistår vesentlig i arbeidet med innlemmelsen og i prinsippet bekoster egen innsats.

11.4 Rettighetshåndtering ved integrering av vokabularer

Alle som bidrar med vokabular som skal innlemmes i Norsk generell tesaurus, inngår en formell avtale med NB. Når vokabularet er avlevert, overtar NB eierskapet til dataene.

12 Veien videre etter NGT 1.0.

Dette kapittelet tar for seg retning og prioriteringer for videre utvikling av NGT etter versjon 1.0, slik denne er spesifisert i kapittel 6.1.

Videre utvikling av NGT (etter NGT 1.0) bør skje langs flere akser for å oppnå et sluttbrukerprodukt med faglig bredde og kvalitet.

- Faglig domene: Det typiske her vil være utbygging av emner som er mangelfullt dekket i NGT 1.0. Det kan imidlertid også være aktuelt å tilpasse faglig dekning den andre veien, dvs. begrense heller enn å utvide, som respons på en eventuell framvekst av nye eller eksisterende emnesystemer på gitte områder, se også 5.5.
- Språklig: Oversettelser til andre språk vil gi flere brukere en inngang til ressursen.
- Knytte forbindelser til et internasjonalt nettverk av emneautoriteter: Mapping/koblinger mot Dewey-klassifikasjon blir viktig for å muliggjøre videre koblinger ut i verden mot andre systemer og flere språk.

12.1 Videre utvikling av faglig domene og begrepsomfang

12.1.1 Utvidelser på områder hvor NGT 1.0 er tynt dekket

Den mest nærliggende måten å gjøre faglige utvidelser på er å identifisere emneområder der NGT bør styrkes faglig og begrepsmessig og hvor det samtidig

finnes et annet vokabular i en norsk institusjon som er interessert i å samarbeide. I slike tilfeller kan det være aktuelt å integrere hele eller deler av dette vokabularet inn i NGT.

Det er foreløpig ikke gjort noen detaljert analyse på emnemessige hull og mangler i NGT, og det er heller ikke hensiktsmessig før de eksisterende vokabularene omtalt i Appendiks 2 er integrert, og NGT 1.0 foreligger. Ut fra det vi kan se om NBs emneordslister samt UBOs faglige profil kan vi likevel anta at for eksempel *ingeniørfag* er mangelfullt representert. Det samme gjelder *idrettsfag*[1].

På disse områdene finnes vokabularene TEKORD[2] og Norsk idrettstesaurus[3] . Begge disse vil være kandidater for å utfylle NGT på hver sine områder og dermed utvide bruksområdet for NGT.

Når det gjelder ingeniørfag, så er dette representert med eget hierarki i Humord, men det er lite utbygd og begreper fra TEKORD vil her kunne styrke NGT. Vokabularet har også 20% overlapp mot Realfagstermer.

Termer fra Norsk idrettstesaurus kan også være aktuelle å integrere i NGT. Tesaurusen er ikke oppdatert siden 1993, men Norges idrettshøgskoles bibliotek har laget en emneordsliste med supplerende termer i ettertid. Den opprinnelige tesaurusen er publisert som tekstdokument, - det vites ikke om den fortsatt er tilgjengelig i elektronisk form.

En forutsetning for å integrere disse vokabularene er selvfølgelig at eierne/emne-miljøene ser seg tjent med en fellesløsning og dermed er interessert i et samarbeid.

12.1.2 Utvidelser som følge av forespørsler utenfra

Norsk musikkbibliotekforening har nylig sendt en formell henvendelse og bedt NB om å overta ansvaret for den todelte tesaurusen Emneord for musikk[4]. Spørsmålet er ikke fullt ut avklart, men hvis NB skal forvalte Emneord for musikk, bør den gjøres interoperabel med NGT på et eller annet nivå, - i første omgang ved at den forvaltes innenfor samme infrastruktur som NGT, i neste omgang bør integrering av selve vokabularet vurderes.

12.1.3 Tilpasning av NGTs faglige dekning til andre vokabularer

På noen områder finnes fagspesifikke emnesystemer som er godt etablerte i norske fag- og forskningsbibliotek. I slike tilfeller vil det være naturlig at NGT tilpasser sin dekning av disse fagområdene og slik unngår aktiv duplisering av termer på området, se også 5.5.

---

[1] I NORART er bare 1,3% av artiklene klassifisert under Sport/idrett (DDK5: 796-799) og 2,6% klassifisert under ingeniørfag (DDK5: 620-629)
[2] http://datahub.io/dataset/tekord
[3] http://www.nih.no/Documents/Bibliotek/Idrettstesaurus.pdf
[4] http://bergenbibliotek.no/musikk/emneord-for-musikk

Emnesystemet MeSH innenfor helsefag og psykologi er allerede identifisert som et slikt system, og NGTs håndtering av disse emnene skal utredes og gjennomføres innenfor rammen av versjon 1.0, se 6.1 og 7 (Aktivitetene 4 og 5). I den grad noe gjenstår av *Aktivitet 5 Utfør vedtak om begreper innenfor helsefag og psykologi*, må dette utføres etter lansering av NGT 1.0.

Etter som flere fagspesifikke vokabularer oppdateres og konverteres til lenkede data, blir de mer tilgjengelige for bruk utenfor sitt opprinnelsesmiljø. Det er derfor ikke utenkelig at noen norske fag- og forskningsbibliotek velger å ta i bruk systemer som i sin natur vil være mer detaljert på sitt område enn NGT kan ta mål av seg til å være. Avhengig av hvor anerkjente og utbredt disse systemene er, må det vurderes om/hvordan NGT skal tilpasse sin faglige dekning på tilsvarende emneområder.

Foreløpig ser vi at AGROVOC[1] (omfatter bl.a. mat og ernæring, jordbruk, skogbruk og fiske og brukes i dag av Norges miljø- og biovitenskapelige universitet) kan tenkes å få en slik posisjon.

12.1.4 Mapping til andre emneordssystemer

Å opprette maskinlesbare forbindelser mellom emneautoriteter gjør det mulig å gjenfinne informasjon på tvers av emnevokabularer. Hvis NGT mappes til et annet vokabular *V*, kan vi bruke NGT til å søke etter informasjon i samlinger som er indeksert ved vokabularet *V*. Virkningen av dette avhenger av hvor mange koblinger det faktisk er mellom NGT og *V*.

Dewey Decimal Classification

Noen emnesystemer har status som nav eller «hubs» ved at de har stor utbredelse, og at det er etablert koblinger fra svært mange andre vokabularer til disse. Dewey er opplagt et slikt nav, ikke minst gjennom WebDewey, som i nær framtid lanserer full versjon av Dewey på norsk som digital tjeneste. Dewey framstår derfor som det viktigste emnesystemet å mappe NGT til. Gjennom dette oppnår man også indirekte, språkuavhengige koblinger til andre emnesystemer, for eksempel LCSH, BIBBI Emner og til en viss grad AGROVOC.

Når det gjelder mapping, har UBO allerede flere aktiviteter uavhengig av NGT, både pågående og under oppstart. Disse er støttet av utviklingsmidler fra NB. Utvikling av metode for mapping mot Dewey har pågått i 2014, mens prosjektet Mapping mot WebDewey (mapping av Humord og Realfagstermer til Dewey) starter opp i 2015. Mye av dette vil kunne tas rett inn i NGT, og gir også en del indirekte mappinger mellom TEKORD og Dewey. Resterende deler av NGT må mappes til Dewey når versjon 1.0. er ferdigstilt.

Andre emnesystemer

Ordnøkkelen[1] er Riksantikvarens vokabular som brukes til å beskrive kulturarvsobjekter, primært bilder (motiv). NBs interne prosjekt *Emner i NB* (Ohren,

[1] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

Rydland et al. 2013) er i ferd med å utrede anbefalinger om emnebeskrivelse av bildemateriale, og i den sammenheng kan for eksempel Ordnøkkelen være aktuell. Avhengig av hvilke anbefalinger som kommer ut av prosjektet, kan det være relevant å mappe NGT mot Ordnøkkelen. Dette og eventuelle andre emnesystemer må derfor vurderes når anbefalingene fra Emner i NB om bildemateriale foreligger.

FIAF[2]s General Subject Headings[3] for litteratur om film ble av Emner I NB anbefalt brukt til NBs samling av filmlitteratur. Det kan derfor være aktuelt å mappe NGT mot denne. En utfordring med FIAF er at den foreløpig bare er publisert som tekst. Mapping mot FIAF bør derfor avventes til den foreligger som åpne, maskinlesbare data (lenkede data).

12.2 NGT på flere språk

I NGT 1.0 vil alle termene foreligge på norsk bokmål, i tillegg til at noen begrep også vil arve engelske termer fra sitt opprinnelsesvokabular, se Appendiks 2 for språklig dekning i hvert vokabular.

12.2.1 Språk i Norge

På sikt er det et mål at så mye som mulig av NGT også skal ha termer på de andre offisielle og anerkjente språkene i Norge, dvs. nynorsk, samisk (i første omgang nordsamisk) og kvensk. Så vidt mulig bør både foretrukne og alternative termer oversettes.

Når det gjelder nynorsk finnes det modne språkteknologiske verktøy for automatisk oversettelse mellom bokmål og nynorsk, konkret er Nyno[4] et slikt verktøy. Realfagstermer oversettes til nynorsk i 2015 med midler tildelt Universitetsbiblioteket i Bergen fra Kulturdepartementet.

På det samiske språkområdet foregår det løpende terminologiutvikling[5], og NGT bør reflektere status på samisk fagterminologi. Dette er også i tråd med Stortingsmeldingen *Mål og meining* (Det kongelege Kultur- og kyrkjedepartement 2007-2008) som framhever betydningen av å kunne bruke samisk i faglige sammenhenger. Inkludering av samiske termer i NGT vil blant annet føre til at det blir mulig å søke i Samisk bibliografi på samisk, noe som ikke er mulig i dag. Et viktig moment er også at det arbeides med å få til en felles nordisk samisk bibliografi. I en slik sammenheng vil samiske språk til en viss grad være språk som er felles på tvers av landegrensene mellom Norge, Sverige, Finland og Russland.

Kvensk ble anerkjent som eget språk i 2005. Riktignok har kvensk til nå vært lite brukt i Norge, muntlig så vel som skriftlig. *Mål og meining* uttrykker imidlertid sterk

---

[1] http://ordnokkelen.ra.no/multites
[2] International Federation of Film Archives/ Fédération Internationale des Archives du Film
[3] http://www.fiafnet.org/uk/publications/iifp_subjectHeadings.html
[4] http://nynodata.no/nn/produkt/nyno
[5] Blant annet er ordbasen www.risten.no et uttrykk for dette. Kan søkes på norsk og flere samiske språk.

politisk vilje til å bevare og utvikle kvensk som kultur- og meningsbærer, og det pågår nå et revitaliseringsarbeid for språket. UiT Norges arktiske universitet tilbyr et årsstudium i kvensk[1], i tillegg finnes et eget kvensk institutt i Porsanger i Finnmark[2], som bl.a. arbeider med kvensk ordbok. På denne bakgrunn er det naturlig at også NGT inkluderer relevante termer i kvensk.

12.2.2 Engelsk

Det er et mål at hele tesaurusen også skal foreligge på engelsk, da spesielt UH-institusjonene har mange brukere med fremmedspråklig bakgrunn. Oversettelse til engelsk vil kunne lettes ved at termer fra den norske Dewey-oversettelsen kan trekkes inn, termer kan hentes fra DB-Pedia, i tillegg kan mappinger mellom Dewey og LCSH være til hjelp. Noe intellektuelt arbeid vil allikevel være nødvendig.

12.2.3 Oversettelsesarbeidet – nye aktører må inn

Nedenfor følger liste av språk i prioritert rekkefølge.

- Engelsk
- Nynorsk
- Nordsamisk
- Andre samiske språk
- Kvensk

Oversettelser vil imidlertid kunne foregå parallelt, da mye av dette må gjøres av ulike miljøer med kjennskap til det aktuelle språket. Oversettelsesarbeidet vil også kunne gjøres uavhengig av annet arbeid med NGT. Til dette arbeidet er det realistisk å skaffe midler fra den delen av virkemiddelapparatet som støtter terminologi- og språkutvikling. Mulige eksterne finansieringskilder for de ulike språkene må derfor kartlegges.

12.3 Prioritering av arbeidet

Som nevnt pågår det allerede noe oversettelsesarbeid og mapping til Dewey. Dette er arbeid som gjøres uavhengig av NGT:

- Oversettelse av Realfagstermer til nynorsk
- Mapping av Humord og Realfagstermer til Dewey

Oversettelser av NGT til andre språk, med engelsk som førsteprioritet, kan gå parallelt med annen utvikling av NGT som integrering av vokabularer og mapping til Dewey.

Prioritering av utviklingsarbeid på NGT etter versjon 1.0:

[1] http://uit.no/om/enhet/artikkel?p_document_id=68146&p_dimension_id=88147
[2] http://www.kvenskinstitutt.no/

Mapping og utvikling av begrepsapparat		Oversettelse til flere språk
Ferdigstille eventuelt gjenstående arbeid fra Aktivitet 5, se kapittel 7.		Engelsk
Fortsette mapping av «ferdige» hierarkier i NGT til Dewey		Nynorsk
Vurdere og eventuelt integrere Norsk musikktesaurus		Nordsamisk
Vurdere og eventuelt integrere TEKORD		Andre samiske språk
Vurdere og eventuelt integrere Norsk Idrettstesaurus		Kvensk
Vurdere og eventuelt mappe til Ordnøkkelen og/eller andre emnesystemer som er i bruk for bildebasert materiale.		
Vurdere og eventuelt mappe til FIAFs General Subject Headings for litteratur om film		
Vurdere og eventuelt tilpasse innholdet i NGT på miljø- og landbruksfeltet til AGROVOC, eventuelt mappe til AGROVOC		

13 Avsluttende kommentarer

For de som kan - og vet å bruke - alle tilgjengelige kanaler, gir det tilgang til et vell av informasjon. Mer enn å finne noe om et emne, er det snakk om å finne den riktige informasjonen.

Vi vet at behovet for å finne informasjon ut fra *emne* er et helt sentralt behov i informasjonsgjenfinning. Vi vet også at å indeksere dokumenter ved hjelp av kontrollerte begrepsapparat er et gode i systemer for gjenfinning. Det krever imidlertid en innsats i forkant for å ta ut en slik gevinst. Å legge til rette for et felles emnesystem som kan brukes av mange bibliotek, og som forvaltes i fellesskap, er en måte å rasjonalisere denne innsatsen.

Også det å knytte en eventuell ny, norsk tesaurus til andre norske og internasjonale systemer er en måte å dra nytte av hverandres arbeidsinnsats. Hvis vi på toppen av dette klarer å utvikle gode brukergrensesnitt og hjelp både ved indeksering og søk, vil vi få et verktøy som gir mye igjen for denne innsatsen.

Hensikten med arbeidet som er beskrevet her, er i siste instans bedre digitale tjenester for brukerne, enten det er snakk om sluttbrukere som selv finner fram til relevant stoff, eller det er bibliotekaren som veileder for sluttbrukeren. Først og fremt vil en norsk generell tesaurus legge et godt *grunnlag* for slike tjenester, ved at kvalitetssikrete emnedata gjøres åpent tilgjengelig for enhver tjenesteleverandør som ønsker å benytte dem, i og utenfor bibliotekfeltet.

Et vellykket resultat av dette prosjektforslaget vil gi oss en verdifull ressurs både kunnskapsorganisatorisk og språklig/terminologisk. Tesaurusen som grunnlag for nyutvikling eller forbedring av søketjenester er kanskje det som er mest nærliggende. Dette omfatter både tjenester som favner kunnskapsfaglig bredt, og spesialtjenester for mer begrensede fagområder, eller i spesielle sammenhenger.

Men også utenfor bibliotekfeltet blir man mer og mer bevisst på verdien av slike manuelt kuraterte ressurser (som en tesaurus jo er), - spesielt innenfor digital humaniora når det gjelder tekstanalyse (text mining). Blant annet viser Språkbankens eksperimentering med emneord og Dewey i tekstanalyse at disse er svært lovende som hjelpemidler til å få mer «semantikk» ut av teksten, - til å «forstå» hva teksten handler om.

Det kan bli en spennende utvikling!

14 Referanser

Det kongelege Kultur- og kyrkjedepartement (2007-2008). Mål og meining. Ein heilskapleg norsk språkpolitikk. Stortingsmelding nr. 35. Oslo.

Hegna, K., M. Almo, et al. (2012). Bibliografisk og emnemessig beskrivelse av UBOs samlinger. Rapport fra en prosjektgruppe. Oslo, Universitetsbiblioteket i Oslo.

Hjortsæter, E. (2009). Emneordskatalogisering. Innholdsanalyse, emnerepresentasjon og lagring. 3. utg. Oslo.

IFLA Working Group on Guidelines for Subject Access by National Bibliographic Agencies (2012). Guidelines for Subject Access in National Bibliograpies. Ed. by Yvonne Jahns.

International Standardization Organization (1986). Documentation : guidelines for the establishment and development of monolingual thesauri. 2nd ed. (ISO 2788). Geneve, ISO.

International Standardization Organization (2011-2013.). Information and documentation: thesauri and interoperability with other vocabularies. Part 1: Thesauri for information retrieval. Part 2: Interoperability with other vocabularies. Geneve, ISO.

Jensen, C. H., Ed. (2014). Kulturstatistikk 2013. Oslo, Statistisk sentralbyrå.

Library of Congress, C. P. a. S. O. (2007). Library of Congress Subject Headings. Pre- vs. Post-Coordination and Related Issues. Washington, DC, Library of Congress.

Nilbe, S. (2012). Semiautomatic merging of two universal thesauri: The case of Estonia. Landry, Patrice, ed. IFLA Series on Bibliographic Control : Subject

Access : Preparing for the Future. Berlin, DEU: Walter de Gruyter, 2011. ProQuest ebrary. Web. 2 January 2015. P. Landry, Walter de Gruyter**:** 51-57.

Ohren, O. P., K. Rydland, et al. (2012). Emneinnganger i Nasjonalbiblioteket; Kartlegging av praksis for emnebeskrivelse. Oslo, Nasjonalbiblioteket.

Ohren, O. P., K. Rydland, et al. (2013). Emneinnganger i Nasjonalbiblioteket; Anbefalinger om praksis for emnebeskrivelse. Del 1: Materiale med verbalt innhold. 03.05.2013. Oslo, Nasjonalbiblioteket.

Sauperl, A. (2009). "Precoordination or not? A new view of the old question." Journal of Documentation **65**(5): 817-833.

APPENDIKS 1: Forslag om forprosjekt

Utvikling av Norske emneord med utgangspunkt i Humord - Forprosjekt

Mål og resultat

For å skaffe innsikt i og oversikt over hva det innebærer å få på plass en universell tesaurus med utgangspunkt i Humord, foreslår vi å gjennomføre et *forprosjekt*, som skal gi et best mulig beslutningsgrunnlag for *om* og eventuelt *hvordan* selve hovedutviklingen bør gjennomføres.

Forprosjektet skal resultere i:

1. Plan for utvikling av Norske emneord versjon 1.0 på bokmål, inkludert aktiviteter, tidsplan med milepæler, deltakere/organisering og ressursestimat.
2. Forslag til driftsmodell for Norske emneord
3. Norske Emneord versjon 0.1 på bokmål.
4. «Veikart» for Norske emneord: Forslag til og plan for videreutvikling. Aktuelle områder for videreutvikling:
    - mapping til andre vokabularer og emnesystemer («vocabulary alignment»)
    - oversettelse til andre språk i Norge (nynorsk, samiske språk, kvensk)
    - eventuell utvidelse av fagområde

Aktiviteter

Arbeidspakke 1: Plan for utvikling av Norske emneord 1.0. {13 uv}

Denne arbeidspakken skal resultere i en mest mulig fullstendig plan for å utvikle Norske emneord 1.0 på bokmål.

Oppgave 1: Overordnet beskrivelse av Norske emneord 1.0 [7 uv]
Avgrensning og omfang av Norske emneord 1.0 (1 uv)

- Definer målgrupper og bruksområde for Norske emneord
- Avgrens tesaurusen faglig og prøv å beskrive omtrentlig faglig nivå
- Hvilke typer begreper skal inkluderes? (bare generelle emneord, ikke sjangre og ikke entiteter som personer, korporasjoner, geografiske navn eller lignende)

Valg av tesaurussystem og annen teknologi (2 uv)

Det er svært viktig å ha et godt og brukervennlig system som støtter utvikling og vedlikehold på en effektiv måte. Aktiviteten innebærer å

- spesifisere krav til systemet. Systemet bør muliggjøre distribuert utvikling og støtte arbeidsflyt.
- skaffe oversikt over relevante alternative systemer
- velge og anskaffe system.

Det bør også vurderes om man trenger annen programvare i tillegg, for eksempel programvare som støtter sammenligning mellom vokabularer, eller programvare som trekker ut sentrale begreper fra tekst.

Spesifiser viktige kilder for begreper og termer (3 uv)

Det ligger fast at Norske emneord tar utgangspunkt i Humord, men andre vokabularer er også aktuelle. Aktiviteten innebærer å:

1. identifisere kilder for nye termer.
    - Eksisterende vokabularer: Nasjonalbibliotekets ulike emneordslister og nøkkelord må inkluderes her. Andre aktuelle er Realfagstermer, TEKORD, Wikipedia-kategorier, DbPedia o.a.
    - Leksikalske språkressurser, som kan være viktige verktøy for å finne korrekte relasjoner mellom begreper, synonymer, oversettelse, med mer. Eksempler: Norsk ordvev for bokmål og nynorsk og fagtermbasen Snorre.
    - Andre kilder: Eksempelvis nye artikler fra sentrale tidsskrifter, fagpersoner, etc.
2. skissere metode for bruk av kildene. Programvare som kan automatisere deler av dette arbeidet blir viktig her.

Spesifiser representasjonsform (0,5 uv)

- Hvordan skal Norske emneord representeres?
- Hvilket format skal det opereres med internt? (Her bør svaret være SKOS i en eller annen variant)
- Hvilke andre formater skal Norske emneord kunne eksporteres til? Minst SKOS, SKOS-XL og MARC/RDF

Spesifiser eventuelle andre krav til Norske emneord (0,5 uv)

Krav som ikke dekkes av det ovenstående, spesifiseres her. Et aktuelt eksempel er krav med hensyn til flerspråklighet.

Oppgave 2. Aktivitetsbeskrivelse og tidsplan [3 uv]

Dette innebærer å beskrive arbeidet som må gjøres for å komme fram til Norske Emneord 1.0.

- Det totale arbeidet brytes ned i arbeidspakker eller aktiviteter, som beskrives kort, inkludert deres innbyrdes avhengigheter.

- Estimér grovt ressursbehov (i uke- eller månedsverk) for hver arbeidspakke/aktivitet.
- Skissér en tidsplan, hvor aktivitetene gis en start- og sluttdato (relativ til oppstartdato) og visualiseres i et Ganttdiagram. Forutsetningene om personellressurser redegjøres for i Oppgave 4.

Oppgave 3: Forslag til organisering av arbeidet med Norske emneord 1.0 [2 uv]

- Utarbeid en hensiktsmessig organisasjonsstruktur: Spesifiser grupper, roller og ansvar.
- Identifiser aktuelle deltakere/institusjoner og hvilke roller de ulike deltakerne skal fylle. Så langt det er mulig, bør vilje til deltakelse avklares med den enkelte institusjon allerede på dette stadiet.
- Tenk gjennom om noe av arbeidet kan/bør settes ut på anbud, eller om personellressurser utenfor deltakergruppen kan/bør innhentes. Dette kan gjelde programvareutvikling eller -tilpasning, behov for spesialkompetanse, IKT-støtte eller annet.

Oppgave 4: Ressursbehov og økonomi [1 uv]

Med utgangspunkt i aktivitets- og tidsplan planen (Oppgave 2) gjøres et estimat over personellbehov i de ulike fasene av utviklingen. Andre kostnader (for eksempel lisenskostnader, innkjøp av maskinvare eller tjenester) identifiseres og estimeres. Finansieringsmodell/kostnadsfordeling mellom deltakerinstitusjonene foreslås.

Arbeidspakke 2: Driftsmodell for Norske emneord {2 uv}

Denne arbeidspakken skal resultere i et forslag til hvordan Norske emneord bør driftes. Beskrivelsen må omfatte følgende:

- Aktører, roller med tilhørende myndighet og ansvar. Her bør det beskrives hvem som har ansvaret for drift og vedlikehold av tesaurussystemet, samt en mulig fagbasert ansvarsfordeling mellom aktørene
- Overordnet beskrivelse av vedlikeholdsprosessen: Arbeidsflyt fra forslag om oppdatering (nytt begrep, endring i begrep, avhend begrep) til godkjent oppdatering.
- Opplegg for tilordning av rettigheter til emneautoritetene må spesifiseres.
- Forslag til finansiering/kostnadsfordeling for drift og vedlikehold Norske emneord
- Spredning/distribusjon:
    - Lisensiering og opphavsrettslige forhold må avklares og beskrives. Norske emneord bør være åpen, og dette må angis eksplisitt i form av brukslisenser.
    - Spesifiser hvilke portaler/ressurs-sider/datasettregistre Norske emneord bør eksponeres i (for eksempel data.norge.no, Termportalen (framtidig Clarino-resultat), andre…
- Forslag til skriftlig avtale mellom driftspartnerne

Arbeidspakke 3: Etablere Norske emneord versjon 0.1 (Humord) {4 uv}

Denne arbeidspakken skal resultere i en initiell (0-te) versjon av Norske emneord.

Det ligger stor verdi i å få fram et praktisk og konkret resultat allerede i forprosjektet, også for å skaffe oss en første erfaring med valgt tilnærming/arbeidsmåte. Vi foreslår derfor å utarbeide en initiell versjon av Norske emneord i forprosjektfasen.

Den initielle versjonen vil i hovedsak tilsvare Humord slik den er nå, dog bare inkludert de *generelle emneordene* (ikke sjangre og geografiske entiteter) og konvertert til vedtatt internformat (SKOS?), og beriket med noen av Nasjonalbibliotekets mindre emneordslister. Hvis mulig, bør tesaurusen realiseres i valgt tesaurussystem, og arbeidsflyten beskrevet i Arbeidspakke 2 bør prøves ut.

Arbeidspakke 4. Veikart for Norske emneord {3 uv}

Resultatet av denne arbeidspakken skal være et forslag til videre utvikling av Norske emneord (etter versjon 1.0), samt en beskrivelse av mulige anvendelser av og tjenester basert på samme.

Oppgave 1. Skissér videre arbeid med Norske emneord (etter versjon 1.0) [1 uv]

- hvilke andre vokabularer tesaurusen bør knyttes an (mappes) mot
- andre språk den bør oversettes til
- eventuell utvidelse av fagområde

Forslagene må prioriteres med hensyn til tid så vel som betydning/viktighet.

Oppgave 3. Mulige anvendelser av Norske emneord [2 uv]

Brukstilfeller (use cases) (0,5 uv)

Det kan være lurt å beskrive noen konkrete, tenkte situasjoner hvor bruk av Norske emneord (eller tjenester basert på Norske emneord) inngår. Brukstilfellene bør innhentes fra brukermiljøene.

Anvendelser og digitale tjenester (1,5 uv)

Foreslå mulige anvendelser av og digitale tjenester basert på Norske emneord. Forslagene kan være på ulikt modenhetsnivå, fra «opplagte» anvendelser (som integrasjon med discoverysystem) til tjenester hvis realisering ligger lenger fram (eksempelvis tjenester som helt eller delvis automatiserer indeksering).

Gjennomføring av forprosjektet

Forprosjektet er aktualisert ved at UBO har søkt om midler til et 3-årig prosjekt for videreutvikling av Humord (bl.a. ved å inkludere Realfagstermer og TEKORD i tesaurusen) og mapping mot WebDewey/DDC23. Prosjektsøknaden stemmer godt

med våre tanker slik de er formulert i Anbefalinger om praksis for emnebeskrivelse. Del 1.

Organisering og deltakere

Det er viktig at Nasjonalbiblioteket tar en aktiv rolle tidlig i dette arbeidet, ikke minst for å sikre at Norske emneord også dekker våre behov. Det er derfor naturlig at Nasjonalbiblioteket og UBO sammen planlegger utviklingen av Norske emneord, dvs. utfører dette forprosjektet i samarbeid. Dette bør avklares med UBO i forbindelse med saksbehandlingen av søknaden.

Ressursbehov og tidsplan

I det ovenstående er det ved hver arbeidspakke/oppgave/deloppgave gitt et estimat for ressursbehov i form av *arbeidstid målt i ukeverk*. Det understrekes at estimatene er meget løselige og grenser til gjetning.

Estimatene tilsier at forprosjektet utgjør en belastning på *22 ukeverk*.

Vi foreslår at forprosjektet bearbeides relativt konsentrert i løpet av første halvår 2014, dvs. ferdigstilles innen *1. juli 2014*.

APPENDIKS 2: Vokabularene som skal inngå i NGT 1.0

Her gis en kort beskrivelse av alle vokabularene som skal inngå i NGT 1.0

Humord

Bakgrunn og bruksområde

Humord er en norsk tesaurus for humaniora og samfunnsvitenskap med tilgrensende fagområder.

Humord er også et indekseringssamarbeid innen rammene av biblioteksystemet BIBSYS.

Humord sprang ut fra et tesaurusprosjekt 1993-1994 der Universitetsbibliotekene i Oslo, Bergen, Tromsø og Trondheim deltok. Grunnlaget var en emneordsliste bygget opp ved Fakultetsbiblioteket HF, Universitetsbiblioteket i Oslo 1988-1994. Prosjektet ble ledet av Norsk termbank, Universitetet i Bergen.

Fra begynnelsen dekket Humord humaniora-fagene, men samfunnsvitenskapelige termer ble tatt gjennom et prosjekt i 2011.

Innhold og omfang

Humord har to typer emneord, samt navn på noen typer entiteter, med andre ord:

- Vanlige innholdsbeskrivende emneord, f.eks. Filosofi, Engelsk språk, Sosialantropologi.
- Formtermer, som beskriver dokumentets bibliografiske eller fysiske form og skiller seg fra de vanlige emneordene ved at de har forklaringen «Form» i parentes. Eksempler: Ordbøker (Form), Tidsskrifter (Form). Formtermene er samlet i et eget delhierarki i Humord.

Navn: Humord inneholder navn på flere typer entiteter, hovedsakelig geografiske steder, tidsperioder, historiske begivenheter, skikkelser fra litteratur, mytologi og religion. Tesaurusen inneholder ikke personnavn og navn på institusjoner. (Slike navn registreres obligatorisk i egne felt i metadata)

Humord har pr. 2015-01-21:

- 18349 hovedtermer
- 8468 se-henvisninger
- Totalt: 26817 termer

Språk: Bokmål

Struktur

Strukturen er i hovedsak hierarkisk med noe bruk av fasetter.

Tesaurusen styres av regelverket i Humord håndbok som er basert på den gamle standarden for tesauri (International Standardization Organization 1986). Systemet er postkoordinert.

Infrastruktur

Humord vedlikeholdes foreløpig via BIBSYS. Når UiO bytter biblioteksystem til Alma i 2016, er det foreløpig uvisst hvor Humord kommer til å vedlikeholdes.

Humord er tilgjengelig fra følgende:

- XML (internt format): http://wgate.bibsys.no/search/pub?base=HUMORD
- Emnesøk mot BIBSYS: http://app.uio.no/ub/emnesok/?id=uhs
- SKOS: http://data.ub.uio.no/dumps/#humord (generelle se-henvisninger er ikke med).
- Autoritetskontroll under katalogisering: Humord er internt tilgjengelig i BIBSYS' katalogiseringsmodul (BIBSYS blåskjerm) som HUME emneregister.

Juridiske emneord og menneskerettighetsvokabularet

Bakgrunn

Juridiske emneord ble utviklet på begynnelsen av 90-tallet, med en alfabetisk og en systematisk del. Den alfabetiske er en autoritetsliste, men fungerte samtidig som register til klasseskjemaet. Den systematiske delen, L-skjemaet, er en del av UBs gamle klassifikasjonssystem der hvert fag var tildelt sin bokstav. Systemet er i bruk ved de juridiske fakultetene i Oslo, Bergen og Tromsø.

Menneskerettighetstermer forvaltes og videreutvikles av Norsk senter for menneskerettigheters (SMR) bibliotek. Menneskerettighetsvokabularet ble opprinnelig utviklet av Human Rights Research and Education Centre of the University of Ottawa, Canada i 1991. SMRs bibliotek har videreutviklet og tilpasset vokabularet, etter at det ble tatt i bruk på slutten av 1990-tallet.

Innhold og omfang

Juridiske emneord er samlingsbasert og inneholder termer som dekker fagområdet rettsvitenskap og tilstøtende områder. Det har innholdsbeskrivende termer og formtermer. Geografiske navn for jurisdiksjon blir også brukt. Årstall/perioder er brukt i mindre omfang, knyttet til historiske fremstillinger.

Menneskerettighetstermer er et kontrollert emneordsvokabular som i hovedsak dekker menneskerettigheter. Emneordene består av innholdsbeskrivende emneord, emneord for form og emneord for sted.

Menneskerettighetstermer inneholder per november 2014 rundt 1200 termer.

Juridiske emneord har ca. 7300 termer, (det er noe usikkert om dette også inkluderer fraser).

Språk: Juridiske termer er på norsk, mens Menneskerettighetstermene er på engelsk.

Struktur

Juridiske emneord bygger på UBs gamle klassifikasjonssystem, L-skjemaet, og hierarkiet ligger i L-skjemaets struktur. Strukturelt følger skjemaet en inndeling i allmenne og spesielle deler. Den allmenne delen omfatter helt generell juridisk litteratur og generell litteratur inndelt etter form. De enkelte fagdelene følger samme struktur, med generelle verk, inndeling basert på form og deretter en struktur som speiler de enkelte fagområdene. Alle termene er knyttet til et L-nummer. Det finnes også noen *se*- og *se også*-henvisninger, men ikke systematisk.

Menneskerettighetstermene har ikke hierarkiske relasjoner per i dag, men har *se også*- og *brukt for*-henvisninger.

Infrastruktur

Juridiske emneord ligger i en intern database. Se http://app.uio.no/ub/ujur/l-skjema/?enkel=0&mode=tre&query=6325

Menneskerettighetstermene er publisert som frie, åpne data, og er tilgjengelige for søk i BIBSYS her:
http://www.ub.uio.no/om/tjenester/emneord/menneskerettighetstermer.html

Realfagstermer

Innhold og omfang

Realfagstermer er et kontrollert, pre-koordinert emneordsvokabular som i hovedsak dekker naturvitenskap, matematikk og informatikk. Følgende emneordstyper er inkludert:

- innholdsbeskrivende emneord
- emneord for form
- emneord for sted
- emneord for tid

Emneordene er publisert som frie, åpne data, og inneholder per januar 2015 rundt 15 000 emneord som kan brukes frittstående, eller kombinert til strenger (nærmere 16 000 strenger per januar 2015).

Språk: Hovedsakelig norsk bokmål, men et prosjekt for å lage en fullstendig nynorsk versjon er under oppstart. I tillegg er drøye 12 % av vokabularet oversatt til engelsk.

Struktur

Realfagstermer er inndelt i fire grupper, en for hver emneordstype. Emnestrenger som konstrueres på basis av de enkeltstående termene inkluderes også i vokabularet. Hvert emneord er uttrykt ved minimum en unik identifikator og en foretrukken term. I tillegg kan emneord ha synonymer, oversettelser og/eller akronymer, klassifikasjonskoder (DDC eller MSC), se også-henvisninger, noter (til internt bruk) og definisjoner (tiltenkt sluttbruker). Realfagstermer har ikke hierarkiske relasjoner per i dag.

Infrastruktur

Realfagstermer er tilgjengelig som RDF/SKOS, med noen utvidelser fra MADS[1]: http://data.ub.uio.no/

Nasjonalbibliotekets emnevokabularer.

Innhold og omfang

NBs emnevokabularer som skal inn i NGT 1.0 omfatter:

- *Emneord brukt i forfatterbibliografiene:* Det er utarbeidet egne lister over emneord for hver av bibliografiene *Bjørnson, Solstad, Hamsun* og *Collett.* Listene er inndelt i sublister etter emnetype: innholdsbeskrivende emneord, sjanger, omtalt person, omtalt verk og omtalt sted. Emneordene er på norsk.
    - Omfang: Unionen av de innholdsbeskrivende termene for alle forfatterbibliografiene utgjør 461 termer, og det er disse som skal inkluderes i NGT 1.0(ikke omfattet geografisk sted og sjanger).
- *Emneord for Samisk bibliografi*: Det er utarbeidet et eget, prekoordinert emnesystem for Samisk bibliografi. Dette foreligger i to versjoner: i) *Kort versjon* som inkluderer emnestrenger bestående av innholdsbeskrivende emneord og underemneord, kvalifikator og form; ii) *lang versjon* der emnestrengene i tillegg kan inkludere geografisk sted.
    - Omfang: Dekomponering av emnestrengene resulterer i totalt 1457 innholdsbeskrivende emneord (inkludert underemneordene)
- *Emneord for Norske og nordiske tidsskriftartikler (NORART):* NORART emneord er en liste over nøkkelord brukt til indeksering av tidsskriftartikler i NORART og således ikke et kontrollert emnesystem i egentlig forstand. Listen brukes imidlertid til kontroll av kandidat-nøkkelord under indeksering.
    - Omfang: Ca. 40000 nøkkelord.

Språk: Alle emneordslistene er på norsk bokmål. Emneordslisten for Samisk bibliografi inneholder imidlertid samiske navn (på steder, verker o.a.).

[1] http://www.loc.gov/standards/mads/

Struktur

Emneordene til Samisk bibliografi er prekoordinert i emnestrenger som kan inneholde emneord, underemneord, kvalifikator, form og geografisk sted. Har også synonymkontroll til en viss grad.

De andre emneordslistene er enkle, flate lister uten henvisninger/synonymer eller emnestrenger.

Infrastruktur

Emneordslistene for forfatterbibliografiene og Samisk bibliografi finnes som tekstfiler eller regneark, og er ikke integrert i katalogiseringsomgivelsen.

NORART emneord er integrert i katalogiseringsomgivelsen og kan eksporteres derfra som en tekstfil.

APPENDIKS 3: Utvikling av pilot (NGT 0.1)

Norsk generell tesaurus 0.1

I forprosjektet har vi gjennomført pilotprosjektet «NGT 0.1» for å høste noen konkrete erfaringer med tesaurusutviklingsarbeid og spesielt med integrasjon av eksisterende emnevokabularer inn i en eksisterende tesaurus. Grunnlaget for piloten har vært Humord konvertert til RDF/SKOS-XL[1], hvori vi har integrert to av NBs emneordlister.

Oppsett av tesaurusprogramvare (VocBench)

Til pilotprosjektet ønsket vi å prøve ut et RDF/SKOS-basert tesaurussystem som var gratis tilgjengelig. Etter en innledende vurdering av tesaurussystemer ble VocBench valgt. I den endelige vurderingen av tesaurussystemer er også kommersielle systemer inkludert, se Appendiks 5..

VocBench er basert på SKOS-XL og datamodellen kan per versjon 2.2 bare i begrenset grad utvides.[2] Via abstraksjonslaget til Sesame-rammeverket støtter systemet ulike RDF-lagre som for eksempel OWLIM og Virtuoso. I piloten har vi testet med OWLIM Lite.

Konvertering av Humord til SKOS-XL

Humord vedlikeholdes per i dag i EMNE-modulen i BIBSYS[3], som er et system for å vedlikeholde énspråklige og termbaserte tesauruser i ISO 2788-tradisjon[4]. Alle termer, både indekstermer og synonymer/henvisninger, har en unik, lokal ID (eks: HUME09767).

I pilotprosjektet har vi konvertert Humord til RDF/SKOS-XL, en modell for flerspråklige og begrepsbaserte kunnskapsorganisasjonssystemer (KOS). Stor utbredelse både i kunnskapsorganisatoriske miljøer og på den semantiske weben generelt gjør RDF/SKOS til en gunstig modell å basere seg på med hensyn til interoperabilitet.

Å basere seg på SKOS innebærer ikke nødvendigvis å begrense seg til SKOS. I pilotprosjektet har vi undersøkt behovet for utvidelser for å representere Humord og en fremtidig NGT på en tilfredsstillende måte. Spesielt utvidelser fra ISO 25964 er aktuelle. De fleste elementene i Humord lot seg mappe til SKOS(-XL) helt uten problemer. Kvalifikatorer representeres som del av termen, ikke i et eget felt, men vi anser ikke det som særlig problematisk. Fasetter og knutetermer vil trenge utvidelser for å kunne brukes på måten de er tiltenkt, men kan reduseres til SKOS ved behov.

---

[1] SKOS <http://www.w3.org/2004/02/skos> med utvidelsen SKOS-XL <http://www.w3.org/2008/05/skos-xl>

[2] Underklasser kan ikke innføres, underegenskaper kan innføres for skos:related og skos:note

[3] Beskrevet på http://www.ub.uio.no/fag/ehylle/14k019572.pdf

[4] ISO 2788 *Guidelines for the establishment and development of monolingual thesauri* av 1986 ble i 2011 erstattet av ISO 25964-1 *Thesauri for information retrieval*.

Det eneste virkelig utfordrende elementet er såkalte generelle se-henvisninger (faktoriseringer), der én term henviser til to begreper – et klart brudd med SKOS-modellen (vanlige se-henvisninger mappes til skos:altLabel). I NGT 0.1 har vi utelatt disse helt, men andre løsninger enn utelatelse er diskutert i Appendiks 4 (om representasjonsmodell).

Integrasjon av eksisterende emnelister

Arbeidet med å integrere vokabularer i NGT for å øke dekningsgraden utover Humord-basisen vil innebære et betydelig arbeid i en eventuell utvikling av NGT 1.0. For å få en forsmak på hva et slikt arbeid kan innebære har vi i pilotprosjektet integrert to av NBs emnelister:

- samisk bibliografi (SAM): Fra Samisk bibliografi er unionen av hovedemneord og underemneord trukket ut, til sammen 1457 emneord. Emneord kodet som stedsnavn eller form er utelatt og strenger er brutt opp slik at hovedemneord og underemneord listes hver for seg.
- forfatterbibliografiene (SBIB): Fra NBs forfatterbibliografier (Hamsun, Bjørnson, Solstad, Undset og Wergeland) er unionen av emneord trukket ut, til sammen 459 emneord.

SBIB er en flat liste av termer, mens SAM har foretrukne og alternative termer (se-henvisninger). Uten definisjoner eller hierarkiske relasjoner å støtte seg, på vil noen termer fremstå som tvetydige siden en term kan ha ulike betydninger (homonymi). I slike tilfeller er den eneste måten å fastsette termens tiltenkte begrepsmessige innhold en sjekk av litteraturbelegget.

Integrasjon i NGT 0.1 har innebåret at hvert emneord i de to emnelistene har blitt (i) innlemmet i et eksisterende emneord, (ii) underordnet et eksisterende emneord eller (iii) forkastet. Ved innlemmelse i eksisterende emneord har kun unike termer blitt beholdt, men i alle tilfeller har det blitt lagt inn en note om termers historikk.

VocBench har per versjon 2.2 ingen funksjonalitet for integrasjon eller *alignment* av vokabularer.[1] Vi har derfor i stedet arbeidet med data i et enkelt tabellformat utenfor VocBench, spesifikt i Excel.

Automatisk prosesseringsarbeid

Tre former for automatisk behandling ble utført for å lette integrasjonsarbeidet:

- Foretrukne termer som hadde eksakt likhet med foretrukne termer i Humord ble automatisk markert for innlemmelse i det tilhørende emneordet. Ingen kontroll for homonymi ble gjort utover rask skanning av listene.
- Foretrukne termer som hadde eksakt likhet med ikke-foretrukne termer i Humord og vice versa ble automatisk markert for intellektuell kontroll.

[1] "Basic support for alignment" er planlagt i en nært forestående versjon.

- Litterære karakterer ble automatisk skilt ut basert på kvalifikator, og ordnet under Humord-begrepet «Navn på fiktive personer og skikkelser». Også i andre vokabularer vil det trolig være slike sett av emneord som enkelt kan klynges i forkant av det intellektuelle vurderingsarbeidet.

Intellektuelt vurderingsarbeid

To personer i Humord-redaksjonen gikk gjennom hver sin liste (SAM og SBIB). Gjennom retningslinjer har de ulike utfallene blitt markert på en standardisert måte slik at resultatene enkelt kunne konverteres til RDF-tripler for import i VocBench i etterkant.

Tabell 1 viser et sammendrag av resultatene. Til sammen lot 63 % av emnene seg innlemme i et eksisterende begrep i Humord, 17 % som underordnet et eksisterende begrep, 17 % lot seg ikke innlemme, mens 3 % ikke lot seg innlemme uten videre diskusjon (status uavklart). Termene som ikke lot seg innlemme inkluderer termer som faller utenfor omfanget til NGT (som geografiske navn og korporasjoner) og flertydige/uklare termer. Termer som gir mening i et lite, avgrenset vokabular gjør ikke nødvendigvis det i et stort og generelt. En del termer kunne blitt innlemmet hvis de hadde blitt endret/klargjort. Noen termer ga også inspirasjon til endringer i Humord. Et integrasjonsarbeid bør derfor ta høyde for at basisen ikke er konstant.

Tabell 1 Resultater fra integrasjonsarbeidet

Innlemmelse	**Metode**	**SAM**	**SBIB**	**Sum**
I eksisterende begrep	Automatisk	721	238	959
	Automatisk foreslått og intellektuelt godkjent	104	26	130
	Manuelt	99[a]	24	123
	Sum	*924 (63 %)*	*288 (63 %)*	*1212 (63 %)*
Underordnet eksisterende begrep	Automatisk	-	56	56
	Manuelt	244	26	270
	Sum	*244 (17 %)*	*82 (18 %)*	*326 (17 %)*
Markert for utelatelse	Manuelt	243[b] (17 %)	84[c] (18 %)	327 (17 %)
Uavklart		46[d] (3 %)	5 (1 %)	51 (3 %)
Sum		1457	459	1916

[a] Hvorav 94 henvisninger og 5 generelle se-henvisninger
[b] Hvorav 7 avslag på maskinelle forslag
[c] Hvorav 2 avslag på maskinelle forslag
[d] Hvorav 8 ble markert med se-også-henvisning, men ingen annen tilordning.

Hastigheten på integrasjonsarbeidet var i snitt 20 emneord i minuttet for SBIB og 53 emneord i minuttet for SAM. Den høye hastigheten skyldes at både frekvensen av vanskelige tilfeller og behovet for å kontakte fagreferenter var lavere enn antatt. Tallene er oppløftende, men overføringsverdien til andre vokabularer er selvfølgelig ukjent, og vi forventer at fagområder som er dårlig utbygd i Humord vil være tyngre

å integrere. I tillegg bør det medregnes tid til at én person til går gjennom listene og kvalitetssikrer, noe som ikke har blitt gjort i dette pilotprosjektet.

Konkrete erfaringer fra det intellektuelle vurderingsarbeidet:

- Alle forslag til integrering av emneord bør sees på av to personer. Dette er en kvalitetssikring av arbeidet som er helt nødvendig.
- Emneordene i NGT 0.1 presenteres med et ID-nummer. Selve termen bør også være synlig for den som skal kontrollere innlemmingen slik at man ikke må bruke tid på oppslag.
- En del sammensatte termer er splittet i NGT (Humord) i henhold til retningslinjene som ble utarbeidet da Humord ble startet opp. I noen tilfeller vil man måtte vurdere om de bør sammenføyes. Et eksempel fra Samisk bibliografi er politisk overvåking. I følge Humord håndbok kan termen splittes fordi hovedleddet (overvåking) er en handling utført på, med eller av attributtet (politikk). Her mener vi utviklingen har gått i retning av å bruke de sammensatte termene i større grad.
- Noen emneordslister dekker smale emneområder og inneholder veldig spesifikke termer. Eksempler på dette finner vi i Samisk bibliografi som inneholder for eksempel: Reinbeite, Reinbyer, Reindriftsanlegg, Reindriftsfag og Reindriftsfond. Dette innebærer større spesifisitet enn det man finner i NGT ellers. Her må det vurderes i hvor stor grad man vil tillate ulikt detaljnivå i hierarkiene.
- En del termer er ikke entydige. For termen adapsjoner fra forfatterbibliografiene har NGT fem ulike betydninger avgrenset med ulike kvalifikatorer (betydningsindikatorer som gjør at man skiller termer som skrives likt, men har ulik betydning). Adapsjoner kan ikke knyttes opp mot en av disse uten nærmere undersøkelse av hva som ligger i termen. Posthistorie fra Samisk bibliografi er og et eksempel på en term som ikke er entydig, et annet er språksenter.
- Adapsjoner er også et eksempel på en term som har ulike skrivemåter. I NGT er Adaptasjoner brukt uten at Adapsjoner var lagt inn som en se-henvisning. Integreringsarbeidet krever dermed en del undersøkelser. I dette tilfellet ble termen tatt inn som et synonym, noe som medførte en forbedring av NGT.
- En del termer kunne ikke uten videre tas inn fordi det manglet en passende overordnet term i Humord. Løsninger i Humord (for eksempel mht. behandling av historiefaget) må revurderes før en kan innlemme for eksempel historietermer.
- Arbeidet ble utført av personer uten spesiell fagfaglig kompetanse. I noen tilfeller var det nødvendig å rådføre seg med fagreferenter. Det må vurderes hvilke personer som skal inkluderes i et slikt integreringsarbeid.

APPENDIKS 4: Representasjonsform

I utgangspunktet er datamodellen for en tesaurus definert av den internasjonale standarden ISO 25964. Denne definerer også et XML-schema som i prinsippet kan brukes til representasjon og utveksling av data. I praksis er det sannsynligvis bare deler av modellen som det vil være aktuelt å implementere, og med tanke på utveksling av data fremstår RDF/SKOS (evt. med tillegg av SKOS-XL) som formatet med mest utbredt støtte. Dette ligger også til rette for publisering av tesaurusen som Linked Data.

SKOS/SKOS-XL kan bare delvis representere datamodellen i ISO 25964. Forholdet mellom SKOS og thesaurus-standarden, med forslag til utvidelser, fremgår av dokumentet *Correspondence between ISO 25964 and SKOS/SKOS-XL models*[1]. Dette inneholder også forslag til utvidelser av SKOS/SKOS-XL som skal favne semantikken i ISO 25964.

I denne omgang har vi forholdt oss til tesaurusen HUMORD slik denne er representert i eget (proprietært) XML-format, og forsøkt å representere denne i SKOS/SKOS-XL med minst mulig tap av informasjon.

Representasjonsform i NGT 0.1

Vi har valgt å bruke SKOS med tillegg av SKOS-XL, delvis for å bevare dato-elementene i HUMORD, i tillegg til eventuelt andre attributter det kan være behov for på term-nivå. HUMORD har (i eksportformatet) en post per term, og attributter som DATO vil være relatert til termene, ikke til begrepene. Bruken av SKOS-XL er i praksis også betinget av at VocBench, som vi har benyttet i NGT 0.1, forutsetter dette som inndateringsformat.

Nedenfor er RDF angitt i turtle-syntaks og med blanke noder for skos-xl:Label for lesbarhetens skyld. (I NGT 0.1 er denne mappingen brukt til inndatering, men med ntriples format istedenfor turtle.)

For hvert humord som ikke er en SE-henvisning (elementet <se-id> i eksportfilen) opprettes et skos:Concept, hvor URI konstrueres på grunnlag av <term-id>.

Eks. (humord fra eksportfilen):

```
<post>
 <term-id>HUME00001</term-id>
 <dato>1994-03-21</dato>
 <hovedemnefrase>Humaniora</hovedemnefrase>
</post>
```

[1] http://www.niso.org/apps/group_public/download.php/12351/Correspondence%20ISO25964-SKOSXL-MADS-2013-12-11.pdf

kan konverteres til SKOS/SKOS-XL:

```
ngt:HUME00001 a skos:Concept ;
  skos:inScheme http://data.nb.no/ngt ;
  skos-xl:prelLabel [
    skos-xl:literalForm "Humaniora"@nb ;
    dct:date "1994-03-21"
  ] ;
  skos:editorialNote "Kilde: HUMORD"@nb .
```

En SE-henvisning gir opphav til en tilføyelse av en altLabel, eks.

```
<post>
 <term-id>HUME00002</term-id>
 <dato>2009-10-22</dato>
 <hovedemnefrase>Humanistiske fag</hovedemnefrase>
 <se-id>HUME00001</se-id>
</post>
```

gir opphav til

```
ngt:HUME00001 skos-xl:altLabel [
  skos-xl:literalForm "Humanistiske fag"@nb ;
  dct:date "2009-10-22"
] .
```

En OT-relasjon (overordnet term, dvs elementene <overordnetterm-id> eller <ox-id>) gjengis som skos:broader, eks.

```
... <overordnetterm-id>HUME00008</overordnetterm-id>
```

gir opphav til

```
... skos:broader ngt:HUME00008
```

En SO (se også) relasjon blir gjengitt som skos:related:

```
... <se-ogsa-id>HUME08415</se-ogsa-id>
```

gir opphav til

```
... skos:related ngt:HUME08415
```

<definisjon> og <noter> gjengis som hhv skos:definition og skos:scopeNote

Elementer i HUMORD uten eksplisitt representasjon i SKOS

Elementene <gen-se-henvisning> og <gen-se-ogsa-henvisning>: Dette svarer til CompundEquivalence i ISO 25964, og er utelatt i NGT 0.1.

Elementet <kvalifikator>: Hvis Humord-posten inneholder et kvalifikator-element føyes teksten til Label i parenteser. Eks.

```
<hovedemnefrase>Gravplasser</hovedemnefrase>
<kvalifikator>Arkeologi</kvalifikator>
```

blir til

```
... skos-xl:literalForm "Gravplasser (Arkeologi)"@nb
```

Elementet <type> i HUMORD har verdiene K eller F (eller er fraværende):

F : fasettindikator : Denne blir ignorert, dvs. termen blir representert som et vanlig begrep. Her vil termen alltid være i parenteser, slik at rollen som fasett fremgår implisitt. Bruken tilsvarer sannsynligvis en ThesaurusArray i ISO 25964.

K : knuteterm. Tilsvarer ThesaurusArray eller ConceptGroup? I tråd med praksis i Humord føyes det til en "#" for å utheve denne i SKOS, eks.:

```
<post>
 <term-id>HUME03954</term-id>
 <dato>1995-12-18</dato>
 <hovedemnefrase>Historie og historiefaget</hovedemnefrase>
 <toppterm-id>HUME03954</toppterm-id>
 <overordnetterm-id>HUME00001</overordnetterm-id>
 <type>K</type>
</post>
```

gir opphav til et Concept:

```
ngt:HUME03954 a skos:Concept ;
  skos:inScheme http://data.nb.no/ngt ;
  skos-xl:prelLabel [
    skos-xl:literalForm "Historie og historiefaget #"@nb ;
    dct:date "1995-12-18"
  ] ;
  skos:broader ngt:HUME00001 ;
  skos:editorialNote "Kilde: HUMORD"@nb .
```

Elementet <toppterm-id> blir ikke eksplisitt representert. Det tilsvarer relasjonen hasTopConcept i ISO 25964 og vil bare være implisitt representert i SKOS.

Noen alternativer

Valgene ovenfor m.h.t. utelatelser og implisitt representasjon av elementer er delvis motivert av hva det valgte verktøyet (VocBench) er i stand til å nyttiggjøre. I

prinsippet kan noen av elementene representeres på alternative måter, f.eks. ved bruk av ISO 25964 SKOS extension[1].

Representasjon av generelle se-henvisninger (faktorisering)

En generell se-henvisning i Humord foreskriver at en sammensatt term skal faktoriseres. Eks:

```
<post>
  <term-id>HUME00247</term-id>
  <hovedemnefrase>Kommunistisk etikk</hovedemnefrase>
  <gen-se-henvisning>Kommunisme * Etikk</gen-se-henvisning>
</post>
```

Denne refererer implisitt til begrepene (forenklet):

```
<post>
  <term-id>HUME00240</term-id>
  <hovedemnefrase>Etikk</hovedemnefrase>
</post>

<post>
  <term-id>HUME09642</term-id>
  <hovedemnefrase>Kommunisme</hovedemnefrase>
</post>
```

Dette kan uttrykkes ved hjelp av skos-thes utvidelsen av skos / skos-xl:

```
ngt:HUME00247 a skos-thes:CompoundEquivalence ;
  skos-thes:plusUF [
    skos-xl:literalForm “Kommunistisk etikk"@nb
  ] ;
  skos-thes:plusUSE ngt:HUME00240label1 ;
  skos-thes:plusUSE ngt:HUME09642label1 .
```

Her er det referanser ikke til begreper men til termer, dvs. instanser av skos-xl:Label, og vi har forutsatt at disse er definert som følger:

```
ngt:HUME00240 skos-xl:prefLabel ngt:HUME00240label1 .
ngt:HUME00240label1 skos-xl:literalForm “Etikk"@nb .
ngt:HUME09642 skos-xl:prefLabel ngt:HUME09642label1 .
ngt:HUME09642label1 skos-xl:literalForm “Kommunisme"@nb .
```

En enklere, alternativ representasjon i SKOS kan være å la henvisningen være et skos:Concept uten prefLabel, for implisitt å markere at det ikke er et “fullverdig” begrep:

[1] http://pub.tenforce.com/schemas/iso25964/skos-thes/

```
ngt:HUME00247 a skos:Concept ;
  skos-xl:altLabel [
    skos-xl:literalForm “Kommunistisk etikk"@nb
  ] ;
  skos:broader ngt:HUME00240 ;
  skos:broader ngt:HUME09642 .
```

Representasjon av fasetter og knutetermer

Emneord merket som fasett eller knuteterm kan alternativt representeres som skos:Collection, med alle underordnede begrep som objekter for skos:member-relasjoner. Eks (noe forenklet for oversiktens skyld):

```
<post>
  <term-id>HUME00005</term-id>
  <hovedemnefrase>Arkeologi</hovedemnefrase>
</post>
…

<post>
  <term-id>HUME00102</term-id>
  <hovedemnefrase>(arkeologi etter type)</hovedemnefrase>
  <overordnetterm-id>HUME00005</overordnetterm-id>
  <type>F</type>
</post>
…

<post>
  <term-id>HUME00103</term-id>
  <hovedemnefrase>Eksperimentell arkeologi</hovedemnefrase>
  <overordnetterm-id>HUME00102</overordnetterm-id>
</post>
…

<post>
  <term-id>HUME00105</term-id>
  <hovedemnefrase>Etnoarkeologi</hovedemnefrase>
  <overordnetterm-id>HUME00102</overordnetterm-id>
</post>
…

<post>
  <term-id>HUME00106</term-id>
  <hovedemnefrase>Historisk arkeologi</hovedemnefrase>
  <overordnetterm-id>HUME00102</overordnetterm-id>
</post>
```

kan gjengis som (forenklet):

```
ngt:HUME00005 a skos:Concept ;
  skos:prefLabel "Arkeologi #"@nb .


ngt:HUME00103 a skos:Concept ;
```

```
  skos:broader ngt:HUME00005 ;
  skos:prefLabel “Eksperimentell arkeologi"@nb .


ngt:HUME00105 a skos:Concept ;
  skos:broader ngt:HUME00005 ;
  skos:prefLabel “Etnoarkeologi"@nb .


ngt:HUME00106 a skos:Concept ;
  skos:broader ngt:HUME00005 ;
  skos:prefLabel “Historisk arkeologi"@nb .


ngt:HUME00102 a skos:Collection ;
  skos:prefLabel “(arkeologi etter type)"@nb ;
  skos:member ngt:HUME00103 ;
  skos:member ngt:HUME00105 ;
  skos:member ngt:HUME00106 .
```

Her forutsettes at rekkefølgen av medlemmer ikke er (semantisk) signifikant, i motsatt fall kan skos:OrderedCollection benyttes.

Merk at en fasett-node (skos:Collection) ikke selv knyttes til begrepshierarkiet med broader/narrower-relasjoner, men alle medlemmene er knyttet til et felles overordnet skos:Concept som (i det originale HUMORD-hierarkiet) er overordnet fasett-noden.

For å kunne knytte fasett-noden til begrepshierarkiet kan den alternativt uttrykkes v.h.a. skos-thes:ThesaurusArray (som er en subClass av skos:Collection) og relateres til overordnet begrep vha. skos-thes:superOrdinate (evt. den inverse skos-thes:subordinateArray). Siste statement ovenfor kan da erstattes med:

```
ngt:HUME00102 a skos-thes:ThesaurusArray ;
  skos-thes:superOrdinate ngt:HUME00005 ;
  skos:prefLabel “(arkeologi etter type)"@nb ;
  skos:member ngt:HUME00103 ;
  skos:member ngt:HUME00105 ;
  skos:member ngt:HUME00106 .
```

Hvis man mangler støtte for Collection eller TesaurusArray kan det også her vurderes et tredje alternativ, å representere fasett-noden som et (ikke fullverdig) begrep i tråd med representasjonen under 1.1, men med bruk av altLabel istedenfor prefLabel:

```
ngt:HUME00102 a skos:Concept ;
  skos:altLabel “(arkeologi etter type)"@nb ;
  skos:narrower ngt:HUME00103 ;
  skos:narrower ngt:HUME00105 ;
  skos:narrower ngt:HUME00106 .
```

Et alternativ som ville gi mer eksplisitt gjengivelse av strukturen i HUMORD er å innføre en dedikerte klasser for fasetter/knutetermer, definert som subklasser av skos:Concept, som i eksempelet kunne gi noe à la:

```
ngt:HUME00102 a ngt-onto:FasettIndikator ;
  skos:prefLabel “(arkeologi etter type)"@nb ;
  skos:narrower ngt:HUME00103 ;
  skos:narrower ngt:HUME00105 ;
  skos:narrower ngt:HUME00106 .
```

APPENDIKS 5: Tesaurussystem for NGT: Anbefaling

Innledning

For å oppnå en effektiv forvaltning av NGT er det viktig å ha et godt system som støtter hele arbeidsprosessen for drift av NGT.

Det finnes en del systemer som retter seg mot forvaltning av tesauri og lignende vokabularer, men ikke svært mange. I stedet for å spesifisere en liste med detaljerte krav til et slikt system fra begynnelsen av, valgte vi derfor å ta utgangspunkt i noen eksisterende systemer som alle virker aktuelle, og sammenligne disse med hensyn på i underkant av 50 parametre, inndelt i følgende grupper:

- Hvordan systemet støtter aktivitetene som inngår i å drifte en tesaurus: Modellering av tesaurusen, innsamling og strukturering av begreper fra ulike kilder, oppdatering av begrepene samt deling av tesaurusen via eksport, publisering av datasett og/eller tilgjengelig endepunkt (f.eks. SPARQL)
- Brukergrensesnitt
- Administrasjon av brukere
- Tilleggsfunksjonalitet
- Ikke-funksjonelle egenskaper

Vurderingen ble gjort ved å fylle ut flest mulig av ovennevnte parametre og samle alt i et regneark[1], som utgjør vår evalueringstabell.

Systemene er valgt ut etter egen kunnskap og ved å konsultere nettressurser med oversikt over slike systemer, for eksempel en oversikt på nettsiden til American Society for Indexing[2].

Til sammen seks systemer ble vurdert:

- MultiTes[3] fra Multisystems, Florida, USA
- Poolparty[4] fra Semantic Web Company (SWC), Wien
- Synaptica[5] fra Synaptica LLC, Colorado, USA
- TemaTres[6] fra R020 Bibliotecología y ciencias de la informacíon, Argentina
- Thesaurus Master®[7] fra DataHarmony, Division of Access Innovations, Inc., New Mexico, USA

[1] https://docs.google.com/spreadsheets/d/1TxIjWINZuXGiy5VDTT0S6aa5l-PjTDCkS5KUMoeJqTk/edit?usp=sheets_home (krever tilgangsautorisasjon)
[2] http://www.asindexing.org/about-indexing/thesauri/thesaurus-management-software/
[3] http://multites.net/index.htm
[4] http://www.poolparty.biz/
[5] http://www.synaptica.com/
[6] http://www.vocabularyserver.com/
[7] http://www.dataharmony.com/services-view/thesaurus-master/

- VocBench[1], utviklet av Food and Agriculture Organization of the United Nations (FAO) og ART Group (Artificial Intelligence Research at Tor Vergata) of the University of Rome 'Tor Vergata'

I det følgende presenteres hvert system kort, basert på evalueringstabellen.

Om de vurderte verktøyene

MultiTes

Dette systemet er utviklet av et firma som har vært på markedet med tesaurusprogramvare siden 1983. Grunnversjonen er MultiTes Pro, en Windowsapplikasjon (enkelt-PC eller Windows server), men systemet tilbys også som «skytjeneste» (MultiTes Online), hvor alle data ligger på Multisystems servere.

Den interne datamodellen er proprietær, men støtter alle de vanlige tesaurusrelasjonene i ANSI/NISO Z39.19-2005, og modellen kan utvides ved at bruker kan definere nye relasjoner. Det skilles ikke mellom begreper og termer, «alt» kalles begreper. Synonymi representeres ved relasjoner mellom begreper, det samme for termer i ulike språk.

Brukergrensesnittet er tradisjonelt Windowsgrensesnitt. Bare MultiTes Online støtter flere brukere.

Flerspråklighet støttes gjennom språkspesifikke relasjoner mellom termer, for eksempel *child FRA enfant*

Interoperabilitet: Tilbyr en rapportgenerator som støtter flere eksportmuligheter, bl.a. eksport til SKOS/RDF og HTML. Kun tesauri formatert som tekst med innrykk kan importeres.

Prising:

- En såkalt «site licence» på Windows Pro m/support koster $4850 (engangs?) + $3600 i support per år.
- MultiRes Online koster $4950 per år. Skytjeneste for publisering av tesaurusen (MultiTes Site) koster $3950 per år

Oppsummert vurdering: MultiTes har isolert sett det vi trenger for tesaurusredigering, men har ingen støtte for andre ting, som f.eks. sammenligning av ulike tesauri. Har også et noe gammelmodig grensesnitt/ utseende

Vurdering basert på: Dokumentasjon og begrenset praktisk utprøving ved prøvelisens på MNultiTes Pro.

[1] http://vocbench.uniroma2.it/

PoolParty Semantic Suite

PoolParty er et relativt omfattende sett av verktøy innenfor det vi kan kalle semantisk web (PoolParty Semantic Suite). Det utvikles stadig (nyeste versjon er datert januar 2015) og inneholder mange produkter/fasiliteter som pakkes på mange ulike måter under ulike navn. PoolParty satser sterkt på avanserte fasiliteter som involverer emnestrukturer på en eller annen måte, alt fra tekstmining/informasjonsekstraksjon, automatisk tagging/indeksering og semantisk søk. For NGT er følgende viktigst:

Kjerneproduktet er PoolParty Thesaurus Server (PPT) som tilbys i fire versjoner/nivåer.

- PoolParty Basic Server: Basisfunksjonalitet for å bygge og vedlikeholde thesauri, med grafisk/tekstlig editor og fullstendig revisjonshistorikk. Det finnes også et begrenset API for å aksessere dataene på en PoolParty server. Den interne datamodellen er mer eller mindre ren SKOS.
- PoolParty Advanced Server: I tillegg til basisfunksjonene finnes mer avanserte fasiliteter, for eksempel
    - Analyse av tekstkorpus både for utvikling av tesaurusen (ekstraksjon av nye kandidattermer) og forslag til indeksering (identifisering av tesaurusbegrep i tekst), ved hjelp av PoolParty Extractor (PPX).
    - Mulighet til å definere sin egen datamodell. Dette må riktignok gjøres ved hjelp av relativt primitive fasiliteter i PoolParty – det er foreløpig ikke mulig å laste inn f.eks. et skjemavokabular beskrevet i et standard språk som OWL, eksempelvis.
    - Støtte for mapping mellom vokabularer, dvs. hovedsakelig redigeringsfunksjoner for lenking mellom begreper i ulike tesauri.
    - Muligheter for beriking av tesaurusen ved lenking til andre datakilder. Følgende datakilder tilbys per default: DbPedia m/kategorier, GeoNames, LCSH og andre datakilder fra LC, Yago og Umbel.
    - Relevante tillegg til ekstra kostnad:
        - Arbeidsflyt for godkjenning av begreper/endringer, basert på distribuert forvaltning av tesaurusen
        - Begrenset støtte for SKOS-XL (SKOS-XL-labler kan genereres)
- PoolParty Enterprise Server: I tillegg til ovenstående tilbys flere funksjoner beregnet på informasjonsforvaltning i virksomheten, eksempelvis automatisk indeksering av innholdssystemer (CMS) som Confluence, Drupal og Sharepoint 2013, sistnevnte til ekstra kostnad.
- PoolParty Semantic Integrator: I tillegg til ovenstående støtter denne semantisk søk. Dette er en kraftig mekanisme som ved hjelp av eksisterende tesauri/taksonomier/kunnskapsgrafer kan konvertere både strukturert og ustrukturert informasjon til rdf, som i sin tur kan søkes i ved hjelp av SPARQL.

For hvert nivå kan mange av tilleggsfunksjonene som er inkludert på et høyere nivå, kjøpes enkeltvis. For NGT vil Advanced Server være den mest aktuelle, men et par ekstra fasiliteter (se ovenfor).

Den interne datamodellen er basert på SKOS, og beriket med elementer fra andre skjemavokabularer som Dublin Core og FOAF.

Flerspråklighet i vokabularene støttes fullt ut

Brukergrensesnitt via nettleser, grafisk og tekstlig. Drag&drop-funksjonalitet støttes.

Interoperabilitet : Import og eksport av ren SKOS støttes. Tilbyr SPARQL-endepunkt i tillegg til en samling APIer.

Teknologi: Standard web-teknologi (Java/Tomcat)

Prising: De ulike versjonene kan kjøpes enten som skytjeneste med månedlig avgift, eller for installering i egen virksomhet, da med en engangs lisens samt årlig avgift for support. Akademisk lisens er mulig for ikke-kommersielle FoU-prosjekter, men ikke som noen varig løsning.

Installering i virksomhet:

- Basic Server: $10,400 (engangsutgift) + $3120 per år
- Advanced Server: $27750 (engangsutgift) + $8,325 pr år
- Enterprise Server: $47100 + $14130 pr år
- Semantic Integrator: $76000 + $22860 pr år
- Skytjeneste for Advanced server koster mellom $1349-$1990 pr måned, etter nivå på support og oppetid bindingstid 1 år.

Oppsummert vurdering: PoolParty er samlet sett et svært kraftig verktøy, og absolutt det mest avanserte av de seks vurderte, i alle fall sett med semantisk web-briller. Funksjonaliteten for selve tesaurusforvaltningen virker imidlertid ikke bedre enn mange andre, det er først og fremst når det gjelder å kombinere språkteknologi/tekstmining med et kontrollert vokabular (både til videreutvikling av vokabularet og til indeksering) at PoolParty utmerker seg. PoolParty er dessuten svært knyttet til SKOS (basisversjonen), fasilitetene for å definere sin egen modell virker ikke veldig brukervennlig.

Vurdering basert på: Dokumentasjon på nettsiden og litt praktisk prøving (demolisens).

Synaptica

Synaptica Enterprise Taxonomy Software (Synaptica KMS) er et tesaurussystem som har sin opprinnelse i 1994, men ser ut til å ha fulgt med i tiden og framstår i dag som

orientert mot semantisk web og linked data. Løsningen selges som et virksomhetsinternt system som støtter databasene Oracle og SQL server.

I tillegg til KMS inneholder Synapticas portefølje to verktøysett som på hver sin måte er komplementære KMS:

- Synaptica Indexing Management System (Synaptica IMS), laget for å støtte manuell indeksering med emnesystemer lagret i Synaptica KMS. Funksjonaliteten er primært beregnet på tagging/indeksering av fulltekst innhold i virksomhetens CMS.
- Synaptica Ontology Publication Suite, som tilbyr et enkelt presentasjonsgrensesnitt for en gitt emnestruktur, enten internt i virksomheten eller på internett

Datamodell: Alle vanlige tesaurusrelasjoner støttes, og den interne modellen kan utvides etter behov både ved å definere nye semantiske relasjoner og nye attributter til de ulike datatypene. Systemet hevder å være kompatibelt med den nye tesaurusstandarden ISO 25954.

Brukergrensesnittet virker intuitivt og lett å bruke, og inkluderer drag&drop-funksjonalitet som hjelpemiddel til å bygge hierarkiene. Har en egen visualiseringsmodul, som gir mulighet for å inspisere og redigere tesaurusen grafisk.

Brukertilgang: Brukere gis rollebasert tilgang, og all redigering skjer i nettleser. Egen støtte for samarbeid i grupper.

Støtter **flerspråklighet** i vokabularene.

Interoperabilitet: Det tilbys et fullt sett med APIer og webservices for interoperabilitet med eksterne applikasjoner. Kan utveksle data på flere formater, - HTML, CSV, RDF og OWL. Det oppgis at både SKOS og SKOS-XL støttes. Det tilbys foreløpig ikke noe SPARQL endepunkt.

Prising: Har en fleksibel prismodell, man kan generelt velge mellom å kjøpe årsabonnement (lisensleie) og å kjøpe en engangslisens koblet ti en årlig avgift for support. Selve prisene er også avhengig av antall brukere, men det er vanskelig å få tall uten å be om tilbud, da priser ikke oppgis på nettsiden. Det oppgis likevel at Synaptica Enterprise KMS starter på $25000, men det er usikkert hvor mye av dette som er en årlig utgift.

Oppsummert vurdering: Synaptica KMS ser ut til å være et fremtidsrettet system som har det vi trenger for utvikling av NGT. Positivt er også støtte for samarbeid i grupper. Ekstrafunksjonalitet som utvikles (f.eks. ved IMS) virker å være orientert mot virksomhetsinternt innhold, og vil trolig ikke være så nyttig for NGT. Det er uvisst hvordan tesaurusen kan presenteres for sluttbruker. Eksemplene på publisering med Synaptica Publication Suite viser bare alfabetisk oppslag. En svakhet er også at

SPARQL ikke støttes. Likevel, Synaptica framstår som et fleksibelt og brukervennlig system for selve tesaurusforvaltningen.

Vurdering basert på: Dokumentasjon på nettsiden og litt epostutveksling med leverandør.

TemaTres

TemaTres controlled vocabulary server er et tesaurussystem utviklet av et lite argentinsk firma; R020 – bibliotecologia y ciencias de la informacion.

Teknologi: TeamTres installeres oppå PHP og krever ellers en webtjener samt database (f.eks. MySQL)

I tillegg til TemaTres tesaurussystem finnes flere komplementære tilleggsverktøy, blant andre:

Thesaurus Web Publisher[1], et verktøy for å publisere et vokabular i TemaTres på web

Tematres Keywords Distiller[2], som ekstraherer nøkkelord fra tekst og kobler disse til termer i en gitt tesaurus. Dette kan tenkes anvendt både som indekseringshjelp (for indeksering av fulltekstdokumenter) og til å identifisere nye kandidattermer innenfor et felt.

Datamodell: De vanlige tesaurusrelasjonene støttes, og modellen kan utvides ved subtyping av disse. I tillegg støttes en fleksibel mengde mappingrelasjoner, som standard alle mappingrelasjonene i SKOS. Det er mulig å merke en term som «metaterm», dvs. at denne ikke skal brukes til indeksering, jfr. Humords «knuteterm». Det skilles ikke mellom begreper og termer, det er altså ikke mulig å tilordne flere termer til samme begrep. Det hevdes likevel (i en bloggpost fra 2012) at TemaTres støtter ISO 25964.

Flerspråklighet støttes kun ved å etablere ett vokabular per språk og lenke termene i de ulike språk/vokabularer ved hjelp av mappingrelasjoner, eksempelvis skos:exactMatch.

Brukertilgang: Svært primitiv brukeradministrasjon, det skilles kun mellom admin og andre. TemaTres støtter derfor ikke en standard arbeidsflyt (fra forslag om endring til godkjenning) på noen god måte.

Interoperabilitet: Eksporterer data til mange formater, bl.a. BS 8723-XML[3], RDF/DC, RDF/SKOS, VDEX, XTM (topic maps), Zthes, JSON og JSON-LD. Tilbyr et SPARQL endepunkt og et API av tilpassede webservices. Import fra SKOS og

[1] http://vocabularyserver.com/thesauruswebpublisher/ (eksempel)
[2] http://vocabularyserver.com/distiller/
[3] BS 8723 (British Standard) er forløper og grunnlag for *ISO 25964*

Prising: TemaTres fri programvare med åpen kildekode, lisensen er GPL 2.0.

Oppsummert vurdering: TemaTres inneholder mye bra funksjonalitet, og isolert sett antakelig det meste av det vi trenger for NGT. Det er likevel en del negative trekk, for eksempel at datamodellen ikke skiller mellom begrep og term, måten å løse flerspråklige tesauri på og mangel på støtte for rollebasert tilgang. Mye av dokumentasjonen på nett er på spansk, og er for en stor del fra 2011. Programvare i seg selv utvikles imidlertid jevnt, nyeste versjon (1.81) var publisert høst 2014. Det er vanskelig å finne informasjon om R020, men siden copyright til TemaTres holdes av en enkeltperson, tyder det på at firmaet er lite (og dermed sårbart). Alt i alt, som produkt betraktet virker hele opplegget litt umodent.

Vurdering basert på: Informasjon på nettsidene samt praktisk utprøving av programvare.

Thesaurus Master®

Thesaurus Master® er et tesaurussystem som utgjør en av tre hovedkomponenter I Data Harmony software suite (MAIstro™), en verktøykasse for informasjonsforvaltning i en organisasjon . De to andre hovedkomponentene i MAIstro™ er :

- Machine-Aided Indexer (M.A.I.™), støtter både automatisk og assistert indeksering av fulltekstdokumenter. Basert på regler (ikke treningssett av dokumenter)
- XML Intranet System (XIS), et XML-basert innholdssystem (CMS)

Teknologi: MAIstro™ forutsetter Java, kjører ellers på mange plattformer, bl.a. Windows og Linux. Kan installeres som enbruker eller i nettverk.

Thesaurus Master® kan også brukes alene («stand-alone»).

Datamodell: Inneholder de vanlige tesaurusrelasjonene i tillegg til ulike note-felter, bl.a. historie og definisjon. Modellen kan utvides med «custom fields». Det hevdes på nettsiden at systemet følger standardene in ISO 25964 and NISO Z39.19.

Flerspråklighet: Dette nevnes ikke eksplisitt noe sted. Men det at systemet følger ISO 25964, forutsetter støtte for flerspråklige tesauri.

Brukertilgang: Brukere kan gis ulike tilganger, og distribuert samarbeid kan gjøres ved hjelp av WebThes®[1]

Interoperabilitet: Thesaurus Master® kan utveksle data på mange formater: SKOS, MARC, Zthes, OWL2, og custom XML.

[1] http://www.accessinn.com:8081/WebViewerStandard/ (eksempel)

Prising: Som vanlig for amerikansk programvare er prisinformasjon lite tilgjengelig uten å be om tilbud, men på et diskusjonsforum fra 2006 oppgis det at nettverksversjonen av Thesaurus Master® alene starter på $25000 (engangsutgift), mens MAIstro™ koster $60000. I tillegg kommer årlig support på 15 % av kjøpesummen.

Oppsummert vurdering: Thesaurus Master® alene virker å være et standard tesaurussystem fra en solid leverandør som sannsynligvis har det meste av basisfunksjonaliteten som NGT trenger, men heller ikke mer. Det nevnes for eksempel ingen støtte for sammenligning/mapping mellom vokabularer. Det mest spennende med DataHarmony-verktøyene er for øvrig indekseringsverktøyet M.A.I.™, men dette blir foreløpig sekundært i NGT-sammenheng. Koblet med en høy pris blir det derfor ikke så relevant.

Vurdering basert på: Informasjon på nettsiden.

VocBench

VocBench er et web-basert verktøy for å redigere/forvalte tesauri og andre typer autoriteter representert som SKOS-XL. Systemet utvikles i et samarbeid mellom Food and Agriculture Organization of the United Nations (FAO)[1] og ART Group[2] ved University of Rome 'Tor Vergata'[3]. Opprinnelig ble VocBench utviklet for intern bruk I FAO, blant annet til drift av AGROVOC. Fra og med versjon 2 ble det imidlertid foretatt et teknologiskifte ved å designe systemet for semantisk web og linked data. Det er også lagt vekt på å designe det slik at bidragsytere skal kunne legge til funksjonalitet som pluggbare komponenter uten av kjernen må røres.

Verktøyet er gjenstand for kontinuerlig utvikling. Løpende versjon er 2.2, - fra versjon 2.3 er det planlagt noe støtte for sammenligning/mapping mellom vokabularer, uten at det er sagt hva denne konkret består i.

For tiden er det flere brukere av VocBench, bl.a. EU Publications office, som forvalter av EUROVOC[4] i VocBench.

Teknologi: Kjører på Tomcat og MySQL. I tillegg anbefales en triplestore, primært GraphDB. EUROVOC har imidlertid satt opp en konfigurasjon med Virtuoso i stedet for GraphDB, noe som er mer aktuelt for NGT.

Datamodell: Standard datamodell er SKOS-XL, men denne kan utvides ved å definere nye egenskaper og relasjoner. (sub-typing). Det er også mulig å laste inn egenspesifiserte datamodeller, slik det er gjort for EuroVoc. Sannsynligvis bør slike være basert på SKOS, dvs. utvidelsene bør gjøres i form av subtyping av SKOS klasser og relasjoner/egenskaper. Det er planer om å utvikle støtte for SKOS-core.

[1] http://www.fao.org/home/en/
[2] http://art.uniroma2.it/
[3] http://web.uniroma2.it/
[4] http://eurovoc.europa.eu/drupal/

Flerspråklighet: Flerspråklige tesauri støttes fullt ut, forutsatt at alle tekststrenger i vokabularet (termer så vel som noter) tilordnes språkkode.

Brukertilgang: VocBench støtter en formalisert arbeidsflyt som er basert på brukernes roller og statusen til de ulike begrepene i tesaurusen (mulige brukerroller og statuser for begreper/termer er predefinerte).

Interoperabilitet: Utveksler data på SKOS-XL, samt tilbyr et SPARQL endepunkt.

Prising: VocBench er fri programvare med åpen kildekode. For de to nevnte triplestore-alternativene (GraphDB og Virtuoso) kan man velge mellom kommersiell og fri versjon. Priser på de kommersielle er ikke undersøkt, men for Virtuoso vil antakelig friversjonen være god nok for NGT.

Oppsummert vurdering: VocBench har de nødvendige støttefunksjonene for å forvalte en tesaurus på en distribuert måte, og med en kontrollert arbeidsflyt. Erfaring viser at det er noen tekniske utfordringer med å bruke gratisversjonen av det anbefalte repositoriet (GraphDB Lite), men vi må regne med at dette bedrer seg etter hvert. Det er også fullt mulig å gjenbruke EUROVOCs konfigurasjon basert på Virtuoso. En negativ ting er at SKOS core ikke støttes, men dette ligger i planene, likeså fasiliteter for sammenligning av vokabularer.

Vurdering basert på: Dokumentasjon, kommunikasjon med utvikler og andre gjennom Google-gruppen vocbench-user og utprøving i NGT v. 0.1

Konklusjon

Slik vi ser det nå, er verken TemaTres eller Thesaurus Master® aktuelle, - førstnevnte på grunn av manglende modenhet og et sårbart utviklingsmiljø, sistnevnte på grunn av høy pris i forhold til funksjonalitet.

Av de andre er både PoolParty, Synaptica og MultiTes aktuelle. Av disse er MultiTes prismessig den rimeligste, men absolutt minst avansert, da er det ikke er mulighet for støtte til sammenligning/lenking mellom vokabularer gjennom verktøyet. PoolParty og Synaptica er begge avanserte, tilbyr mange fasiliteter for lenkede data og kan gi oss mye, men til en høy pris.

VocBench er derimot fri programvare, som vi nå til en viss grad har prøvd ut i NGT 0.1 (piloten) og som vi ser har potensiale til å bli bra. Her har vi også mulighet til å påvirke utviklingen og om ønskelig utvikle egne tilleggsmoduler. I forhold til et rent kommersielt verktøy med betalt vedlikehold, er det imidlertid mer krevende å ta i bruk fri programvare som VocBench, da installering (av VocBench så vel som nødvendige tilleggsprogramvare), konfigurering og generelt oppsett må gjøres av den enkelte brukerinstitusjon. Bruk av VocBench forutsetter derfor at vi har teknisk kyndige personer i prosjektet under hele utviklingen, selv om det er god støtte å finne gjennom brukergruppen vocbench-user og konkret gjennom vår kontakt med EuroVoc-miljøet.

Alt tatt i betraktning, anbefaler vi å ta i bruk VocBench sammen med en triple-store som passer inn i NBs øvrige driftsmiljø.

APPENDIKS 6: Mulige tjenester og bruksområder for NGT

Her skisseres potensielle bruksområder og anvendelser av NGT. Hvilke brukergrupper kan NGT først og fremst være nyttig for - og på hvilken måte? Hvilke tjenester og funksjoner vil NGT kunne bidra med?

Det følgende er hovedsakelig resultat av en intern idédugnad i prosjektgruppa.

Kategorier av brukere

Prosjektet har identifisert 3 hovedtyper av mulige brukere av NGT, men utelukker ikke at andre typer brukere vil finne tesaurusen interessant.

Sluttbrukeren

Med sluttbruker menes en person som forsøker å få svar på sitt informasjonsbehov ved å ta i bruk NGT. En sluttbruker kan være en student, forsker, journalist, privatperson osv. Sluttbrukeren kan ha et annet morsmål enn norsk.

Indeksereren

Dette er bibliotekaren eller fagreferenten som tilordner emneord på dokumenter som skal innlemmes i samlingen – muligens i kombinasjon med katalogisering og/eller klassifisering.

Informasjonsveilederen

Dette kan være en bibliotekar eller fagreferent som veileder en sluttbruker med litteratursøk/faglig orientering m.m.

Digitale tjenester og funksjoner

Resultatet av å gjennomføre prosjektet beskrevet i denne rapporten blir NGT 1.0 som vil være tilgjengelig på nett både som nedlastbart datasett og som navigeringsgrensesnitt for brukeren. Sistnevnte vil i praksis være det grensesnittet som det valgte tesaurussystemet tilbyr eksterne brukere.

Mange spennende og innovative digitale tjenester kan utvikles basert på disse dataene. Det er i stor grad opp til tredjepart/systemleverandører å gjøre bruk av dataene til å utvikle nye tjenester til sin portefølje.

Nedenfor skisseres noen tenkte digitale tjenester basert på NGT- og ønskelige egenskaper ved disse. De to første delkapitlene beskriver tjenester som er relevante i bibliotek.

NGT integrert i bibliotekets discovery-system/søkegrensesnitt

Navigering i emner

NGT bør være tilgjengelig gjennom et lettfattelig brukergrensesnitt med mulighet for enkelt å navigere/browse i emnehierarkier slik at det er mulig å både utvide og

spesifisere søk innen aktuelle emner. På alle nivåer må det være mulig å utføre et søk, vurdere det aktuelle resultatet (trefflista) og deretter å gå tilbake til det aktuelle emnehierarkiet for å justere søket. Synonymer, henvisninger og definisjoner hjelper brukeren med å forbedre/tilpasse sitt søk ytterligere.

Bla og søk på valgt språk

I navigeringsgrensesnittet bør det være mulig å velge språk, dvs. at begrepene i NGT vises med termer på valgt språk (eller et standardspråk der term på valgt språk ikke finnes) Brukeren skal også kunne søke på de ulike språk som NGT tilbyr og få treff på emner som leder til relevant litteratur og ressurser uavhengig av valgte søkespråk.

Direkte fra NGT til informasjonsressursene

Emner skal være knyttet til bestand (litteratur) i biblioteker (trykt og digital) eller lede til ressurser i eksterne kilder/baser utenfor bibliotekene via videresøk, slik at brukeren kan lokalisere ressursen hun/han ønsker. Brukeren skal kunne bestille ønsket litteratur som ikke er tilgjengelig direkte digitalt.

Navigering på tvers av emnesystemer

På sikt vil NGT kunne inneholde mappinger til flere andre emnesystemer, ikke bare Dewey. Ikke alle disse er interessante for alle brukere, og det bør være mulig å slå av og på hvilke som skal inkluderes i det enkelte tilfelle, på en enkel og oversiktlig måte. Når ett eller flere tilmappede emnesystemer er valgt, bør brukeren kunne navigere sømløst mellom begreper i NGT + alle valgte som en helhet.

Hvis sluttbrukeren har behov for søking/navigering i fagområder der litteraturen i stor grad er indeksert ved andre emnesystemer som NGT er mappet til (f.eks. MeSH), bør bruker kunne følge disse mappingene sømløst uten å tenke på eller oppdage at «nå er jeg i MeSH».

Dette er en avansert fasilitet, og systemet bør selvfølgelig settes opp med en standard, som kan være forskjellig for de ulike bibliotek.

Formulering av søk – ikke bare ut fra NGT

Det bør legges vekt på å tilby en god søkeformuleringstjeneste på basis av NGT og andre autoritetsregistre (f.eks. Personer, korporasjoner, geografiske steder, sjanger etc.), ved at

- bruker kan plukke og velge fra NGTs trestruktur, med full oversikt over hierarkiene, samt velge
    - om hierarkiet *under* et begrep skal med i søket
    - stien opp til toppen (de mer generelle begrepene) skal med i søket
- kunne kombinere begreper fra NGT med begreper fra andre autoritetsregistre

I tillegg bør bruker kunne forfine NGT-emnesøk ved hjelp av andre søkemåter, f.eks. at en kan bearbeide resultatet fra et emnesøk med søk i fulltekst i de dokumentene der dokumentet er tilrettelagt for det.

NGT integrert i biblioteksystem

Som støtte til manuell indeksering er det vesentlig at NGT integreres i biblioteksystemet på en slik måte at indekserer til enhver tid har tilgjengelig et godt navigering- og søkegrensesnitt mot tesaurusen. Det er viktig med god oversikt på makronivå samtidig som det er god tilgang til detaljene om det begrep man til enhver tid ser på.

Indekserer skal kunne søke etter kandidatbegreper, se deres plassering i hierarkiet og om nødvendig få tilgang til andre informasjonsressurser som er indeksert med begrepene under vurdering, enkeltvis og i kombinasjon.

NGT kan også brukes i språkteknologisk øyemed

NGT slik den er tenkt utviklet fra og med NGT 1.0 vil i praksis utgjøre en kuratert og kvalitetssikret fagterminologi.

Såkalt emnemodellering (topic modeling[1]) er i dag et svært aktuelt begrep innenfor digital humaniora. Dette er en statistisk metode for text mining[2]/tekstanalyse som går ut på å identifisere *mønstre* i form av grupper av ord som representerer et «emne». Det er nødvendig å ha store tekstkorpus for å få et godt resultat. En utfordring med metoden har vært at det er vanskelig å finne gode betegnelser på emnene. For eksempel kan ordkombinasjonen «smør, gård, avlinger» identifiseres som et emne, men hva skal emnet kalles? Her kan NGT være et godt bidrag. NB som forvalter store mengder fulltekst på norsk eksperimenterer generelt med hvordan emnestrukturer kan utnyttes i tekstanalyse.

NGT og sluttbrukeren

NGT må være tilrettelagt med en sluttbrukertjeneste som er enkelt tilgjengelig, og i størst mulig grad selvinstruende også for brukere som ikke har lang erfaring med denne typen søketjenester.

Nedenfor listes eksempler på hva tesaurusen vil kunne hjelpe sluttbrukeren med.

- Flerspråklig søk og navigering i emner/faglige hierarkier
- Gjøre seg kjent med fagterminologi innen et tema
- Ønske om å orientere seg innen et emne- eller fagområde
- Bestille litteratur innen et emne- eller fagområde for f.eks. et skriftlig arbeid
- Finne stoff/ressurser innen et emne uavhengig av fysisk form og lokalisering
- Lage litteraturliste om et emne
- Finne definisjoner på faguttrykk
- Oversette faguttrykk mellom flere språk

[1] http://en.wikipedia.org/wiki/Topic_model og http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/
[2] Norsk betegnelse mangler

- Hjelp med emnestrukturering og hierarkier i forbindelse med f.eks. arbeid med en ny nettside
- Oppdage nye og interessante emner/tema ved browsing
- Tilby mulighet for høy spesifisitet i søk, samtidig som det er tilrettelagt for browsing.

NGT som støtte til emneindeksering og klassifikasjon

Bibliotekar/indekserer bruker NGT til emnebeskrivelse av dokumenter og ressurser i samlingen, sannsynligvis i et eget indekseringsverktøy for dette formålet. Bibliotekaren kan videre søke og browse i relevante emner for å finne riktig Deweynummer til klassifikasjon. I indekseringsverktøyet kan valgte emneord og klassifikasjon kobles til katalogposten i biblioteksystemet. Hvis emnet befinner seg i andre selvstendige emnesystemer utenfor NGT (MeSH/AGROVOC), vil søket gå mot disse via samme brukergrensesnitt.

Automatisk eller halv-automatisk indeksering

Vi har en visjon om at NGT skal kunne brukes til automatisk eller systemstøttet indeksering av fulltekstdokumenter. En tilnærming til dette kan være fagområdet *maskinlæring*[1], hvor det utvikles metoder som gjør at dataprogrammer kan lære av data.

Etter hvert som NGT tas i bruk, vil det vokse fram en stor base av NGT-indekserte fulltekstdokumenter hvor det kan antas at kvaliteten på indekseringen er høy. Ved å bruke maskinlæringsmetoder og manuelt indekserte dokumenter som «treningsdokumenter» kan vi tenke oss en tjeneste som automatisk foreslår begreper fra NGT for tilsvarende dokumenter som ikke er indeksert. En del av de dyreste tesaurussystemene undersøkt i Appendiks 5 inneholder fasiliteter for automatisk, tesaurusbasert indeksering.

NGT og referansebibliotekaren/veilederen

Referansebibliotekaren kan utnytte tesaurusen på samme måte som sluttbrukeren (se ovenfor) i f.eks. en veiledningssituasjon. I tillegg har referansebibliotekaren tilgang til et grensesnitt for superbrukere som gir tilleggsmuligheter for mer avansert søk og gjenfinning ved f.eks. å vise mappinger mellom emneord og Dewey og gi søke-/browsemuligheter i Dewey. Bibliotekaren kan altså bygge søk ved å bruke både emneordshierarkiet og mappet klassifikasjon.

[1] http://en.wikipedia.org/wiki/Machine_learning

APPENDIKS 7: Språkteknologiske metoder i tesaurusbygging

Språkteknologi kan bidra i forskjellige faser av tesaurusbyggingen for NGT. Fra preprosessering av eksisterende emneord på formnivå, til konstruksjon og evaluering av betydningsrelasjoner.

Formnivå

Samme begrep kan skrives på forskjellige måter, og her vil en analyse av skriftvarianter og bøyningsmorfologi være relevant. For eksempel så vil temaet *elv* i BIBSYS være skrevet som *Elver, elver, elv, elven.* Foruten variasjon mellom stor og liten forbokstav, er det morfologisk varianter som ubestemt form flertall, grunnform og bestemt form entall.

Det mønsteret går igjen på flere emneord. Ordformer kan kobles opp mot Språkrådets fullformsordlister for opprydding og gruppering av termer. I tilfellet med begrepet *elv* kunne man tenke seg å gruppere formene sammen til en eneste mengde som {*elven, elver, elv*}. Merk at hver form også kan være kapitalisert.

I tillegg til bøyningsvarianter kommer skrivefeil, og selv om de er sjeldne kan det være nyttig å ha metoder for å finne dem og koble dem til rett term.

Foruten varianter i skrivemåter inngår emneordene i forskjellig konstruksjoner, som koordinasjoner i *kvinner og samfunn,* samt spesifikke konstruksjoner gjeldende for enkelte emneordsystem (emne-modifikator konstruksjoner). Konstruksjonene er gjerne underlagt forskjellige konvensjoner, som *kvinner : arbeidsliv, kvinner i arbeidslivet,* og *kvinner. historie.* Bruken av punktum og kolon varierer noe, og overlapper i betydning. Her kan man benytte metoder for parsing av naturlig språk for å finne klasser av konstruksjoner for gruppering og konstruksjonenes interne struktur. Emner kan kobles sammen på tvers av slike konvensjoner basert på semantikken til konvensjonene.

Som en del av preprosesseringen og integrering av emneordssystem vil NGT profitere på en språklig analyse av form og struktur, med bruk av språkteknologiske analyseverktøy (formelle grammatikker, og automater knyttet til dem) benyttes for systematisering, gruppering og mapping mellom emneordsystemer.

Betydningsnivå

Med betydningsnivå forstås hvordan ord forholder seg semantisk til hverandre. Det er særlig tre akser som kan studeres:

- taksonomiske hierarkier som forholdet mellom *robåt* og *båt,*
- del-hele forhold som forholdet mellom *kjøl* og *båt,*
- nettverk bundet sammen av relasjoner som forholdet mellom *sjø* og *båt* (en *båt* brukes på *sjø*en).

Betydningsnivåene kan brukes til å sammenligne emneord seg imellom eller sammenligne emneord for et verk med innholdsordene i det[1]. Med den rike tilgangen til digitale tekster vil det siste være en reell mulighet for strukturering av emneord langs forskjellige semantiske akser. Det er flere ressurser som kan benyttes for akkurat det formålet.

Norsk Ordvev inneholder taksonomisk informasjon i tillegg til relasjonell informasjon om ca. 190 000 ord i norsk bokmål og nynorsk (tilgjengelig fra Språkbanken). Ordveven stammer i store trekk fra den danske versjonen av WordNet,

I Ordveven er alle ordene oppgitt med grunnform, og kan kobles til emneordene gjennom en morfologisk analyse som beskrevet ovenfor. Taksonomien i Ordveven kan så benyttes til å strukturere emneord i et betydningshierarki, i tillegg til å finne semantiske relasjoner mellom emneord.

Statistiske metoder

Metoder fra statistisk analyse av språkdata, som assosiasjonsmål for kollokasjoner passer godt for å sammenligne emneord med DDK og innholdsord. NBs database over digitaliserte tekster (for bokmål utgjør de over 220 000 tekster tilgjengelig i digital form) muliggjør en analyse av emneord gjennom måten de forholder seg til innholdsord[2].

Statistiske assosiasjonsmål danner grunnlaget for å foreta en analyse ved hjelp av FCA (Formal Concept Analysis[3]). FCA er en metode for å bygge taksonomiske hierarkier basert på en subjekt-predikat-relasjon. For eksempel kan man tenke seg innholdsordene i en tekst som subjekter og emneordene som predikater for innholdsordene. Assosiasjonsmål mellom innholds- og emneord vil plukke ut delmengder av dem, delmengder som kan benyttes i videre oppbygging. Eller, man kan la desimalkoder i Dewey fungere som predikater og la emneordene være subjekter.

Vi illustreres med et eksempel fra NORART sitt emneordsystem, og lar emneordene være subjekter for desimalkoder. Emneord og kode vil da forme ordnede par bestående av et emneord pluss en sifferkombinasjon fra Dewey.

Flere emneord kan falle inn under samme klassifikasjon, og ett emneord kan kobles sammen med andre emneord på forskjellige måter. Forskjeller i gruppering kan speile betydningsvarianter og innsnevring av betydning. For eksempel vil emneordet *kaffe* forekomme sammen med *vin, øl* og *brennevin* under klassifikasjonen 641.2 (Drikkevarer), mens *kaffe* forekommer sammen med blant andre *matproduksjon* og

[1] For eksempel Lars G Johnsen "Digitalisering og klassifikasjon" i Bibliotecha Nova 4-2014.
[2] Lars G. Johnsen (ibid)
[3] http://en.wikipedia.org/wiki/Formal_concept_analysis

utviklingsland under 338.17 (Produkter). Den første gruppen viser *kaffe* i betydningen drikk, den andre om dyrking og foredling av kaffebønner.

For å systematisere denne krysskoblingen vil man i FCA ta utgangspunkt i mengden av slike par mellom emner og desimalkoder: NORART+Dewey. Konstruksjonen er automatisk, og benytter standard algoritme fra grunnlagsdata[1].

Et eksempel på et par i slik mengde er {*kaffe*, vin} + {641.2, 613,2} (613.2 - Diettetikk), som danner et underbegrep av drikker. Slike mengdepar utgjør de formelle begrepene, og er relatert til hverandre gjennom delmengderelasjonen. Den relasjonen gir så opphav til en ordning som svarer til en taksonomi over emneordene, samt en taksonomi over desimalklassene.

Hierarkiet gitt av ordningen utgjør ikke et tre med klare partisjoner mellom kategoriene. Resultatet er et mange-til-mange hierarki som kan benyttes til tolkning av bruksmåter av emneord (også desimalklassifikasjoner).

Metoden vil kunne belyse forskjellige forslag til taksonomisk hierarkisering basert på statistisk materiale, enten fra emneord + Dewey, eller emneord + innholdsord. Den samme metoden kan benyttes på begge datasettene. Koblet sammen med kvalitative ordboksdatabaser åpner det seg nye muligheter for å studere emneordenes betydning og forhold til hverandre, samt emneordenes forhold til de verk som de beskriver.

[1] Se e.g. Deng, Hiao, Wang (2011) Formal Concept Analysis Based on Rough Set Theory and a Construction Algorithm of Rough Concept Lattice,Springer Berlin Heidelberg

APPENDIKS 8: Oversikt over et utvalg generelle emneordsystemer fra ulike land

«IFLAs Guidelines for Subject Access in National Bibliographies" (IFLA Working Group on Guidelines for Subject Access by National Bibliographic Agencies 2012) er I stor grad brukt som kilde. Utkastet til retningslinjene (Draft, 2011) finnes I fulltekst her: http://www.ifla.org/files/assets/classification-and-indexing/subject-access-by-national-bibliographic-agencies/nba_guidelines_draft_2011-05.pdf.

Det har vært vanskelig å finne mye om hvordan arbeidet egentlig er organisert, men et generelt inntrykk er at nasjonalbibliotekene spiller en viktig rolle.

- **LCSH (Library of Congress Subject Headings)**
  http://www.loc.gov/aba/cataloging/subject/

  Bakgrunn: LCSH danner mønster for mange nasjonale emneordssystemer. Systemet er dominerende i engelsktalende land, men er også oversatt og brukt bl.a. i flere spansktalende land og som utgangspunkt for f.eks. Svenska ämnesord.

  Struktur: Har prekoordinerte emneord i strenger, homonym- og synonymkontroll samt se- og se-også-henvisninger.

  Omfang: 5 millioner emneautoriteter.

  Format: Marc 21, SKOS.

  Tilrettelegging for semantisk web: http://id.loc.gov/

  Mapping til Dewey: Ja.
  Organisering: Utviklet og vedlikeholdt av Library of Congress.

- **Noen land og/eller nasjonalbibliografier som bruker eller tar utgangspunkt i LCSH:**

  - **ANBD (Australian National Bibliographic Database)**
    http://www.nla.au/services/standards.html#SubjectCat
    Bakgrunn: ANBD er LCSH med egne utvidelser for australske forhold. Gjennom prosjektet Australian Subject Access Project arbeides det for å dekke termer om australske forhold (http://www.nla.gov.au/librariesaustralia/slash.html)

  - **Canadian Subject Headings**
    http://www.bac-lac.gc.ca/eng/services/canadian-subject-headings/Pages/canadian-subject-headings.aspx
    Bakgrunn: Canadiana (den kanadiske nasjonalbibliografien) bruker LCSH, men med Canadian Subject Headings som supplement. Finnes bare på engelsk, men har motsvarende franske termer I Répertoire de vedettes-matiere (RVM). RVM inkluderer også nesten hele LCSH.

  - **Andre emneordssytemeri som helt eller delvis bygger på LCSH: Chile, Latvia, Litauen, Namibia, Sør-Afrika.**
    Kilde: Subject access in national bibliographies (IFLA, 2011)

- **FAST (Faceted Application of Subject Terminology)**
  http://www.oclc.org/research/activities/fast/default.htm

  Bakgrunn: FAST tar utgangspunkt i LCSH, og utgangspunktet var et behov for et enklere emneordssystem enn LCSH. Hensikten med å utvikle FAST er at det skal være enkelt for brukerne, og enkelt for sluttbrukerne. FAST brukes bl.a. av Nasjonalbiblioteket på New Zealand og RMIT (teknisk universitet, Australia) til artikkelmetadata. Den brukes i tillegg av flere andre bibliotek særlig til digitale ressurser.

  Struktur: Vokabularet er tatt fra LCSH, men det er delvis fasettert. Dette betyr at mange emneord som i LCSH er i strenger, er delt opp i enkeltord i FAST. Systemet er delt inn i åtte fasetter: Personnavn, korporasjoner, geografiske navn, hendelser, (Standard)titler, tidsperioder, generelle emner (topics) og form/sjanger. Har homonym- og synonymkontoroll, samt se- og se-også-henvisninger.

  Omfang: FAST har mer enn 1 700 000 autoriteter. Dette utgjør mange færre enn LCSH fordi det er enkeltemneord.
  http://www.oclc.org/research/activities/fast/download.html

  Organisering: Utviklet og vedlikeholdt av OCLC.

- **Svenska ämnesord**
  http://www.kb.se/katalogisering/Svenska-amnesord/

  Bakgrunn: Svenska ämnesord er utviklet av Kungliga Biblioteket og brukes i svenske fagbibliotek og i KB (nasjonalbiblioteket). Det omfatter alle emneområder, men brukes hovedsakelig innen et bredt humanistisk og samfunnsvitenskapelig emneområde. Redaksjonen er ved Kungliga Biblioteket. Systemet ble utviklet ved at emneord brukt i svenske fagbibliotek ble samlet og ”vasket”. Det ble tatt i bruk i 2000. De mest brukte termene fra LCSH ble oversatt til svensk. Nye emneord blir kontinuerlig mappet til eksisterende LCSH-termer og til Dewey, fram til 2011 også til det svenske klassifikasjonssystemet SAB.

  Svenska ämnesord består av tre ulike emneordlister: Svenska ämnesord (SAO), Teasurus för grafiskt materila (TGM) og Barnämneord (Barn). Det er gjennomført en forenkling av systemet, slik at det har nærmest seg FAST-struktur.

  Se her for mer informasjon:
  https://www.youtube.com/watch?v=yqihZsHLWaQ&feature=youtu.be

  Struktur: Har prekoordinerte strenger. Forenkling og fasettering (8 fasetter).

  Omfang: 38000 termer

  Tilrettelegging for semantisk web: Det finnes planer, og Svenska ämnesord blir mappet til Dewey ved hjelp av SKOS.

  Organisering: Kungliga Biblioteket har redaktøransvaret.

- **Finto**
  http://finto.fi/en/

  Bakgrunn: Finto er en finsk service for publisering og bruk av vokabularer, ontologier og klassifikasjon. Det er utviklet som et samarbeidsprosjekt mellom det finske Nasjonalbiblioteket, Finansministeriet og Utdannings- og kulturministeriet (ONKI http://onki.fi/).

  Både allmenne tesauruser og spesialvokabularer er gjort tilgjengelig gjennom Finto, ikke bare bibliotekressurser.

  Organisering: Organisert som et felles prosjekt, ONKI, mellom Nasjonalbiblioteket, Finansministeriet og Utdannings- og kulturministeriet.

- **Schlagwortnormtdatei (SWD)**
  http://de.wikipedia.org/wiki/Schlagwortnormdatei

  Bakgrunn: SWD er emnesystemet som blir brukt i Tyskland, Sveits og Østerrike. Det er laget på grunnlag av ordtilfang fra emne-systemer i tyske bibliotek.
  Struktur: Tesaurusstruktur

  Omfang: Ca 550 000 termer

  Tilrettelegging for semantisk web: Emneordene er tilrettelagt for semantisk web som en del av Linked Data-prosjektet ved det tyske nasjonalbiblioteket. Systemet er slått sammen med navne- og korporasjonsautoriteter til Gemeinsame Normdatei (GNB http://openbiblio.net/2012/01/26/german-national-library-goes-lod-publishes-national-bibliography/) . Det er mappet til andre tesauri, og til Rameau og LCSH.
  Mapping til Dewey: SWD er mappet til Dewey gjennom CrissCross-prosjektet: http://linux2.fbi.fh-koeln.de/crisscross/swd-ddc-mapping_en.html

  Organisering: Driftes av det tyske nasjonalbiblioteket i samarbeid med bibliotekene.

- **RAMEAU (Répertoire d'autorité-matiére encyclopédique et alphabétique)**
  http://rameau.bnf.fr/

  Bakgrunn: RAMEAU brukes bl.a. av Bibliothèque nationale de France, franske universitetsbibliotek, samt noen forskningsbibliotek og folkebibliotek. RAMEAU kommer opprinnelig fra Laval RVM(?) og fra LCSH.
  Omfang: 2560 000 termer, hvorav 88 000 er allmenne og 46 000 geografiske.

  Tilrettelegging for semantisk web: Emneordene er tilrettelagt for semantisk web ved hjelp av SKOS. Den systematiske indekseringsdelen er organisert etter hierarkiene I DDC http://www.cs.vu.nl/STITCH/rameau/ (Guidelines for Subject Access in National Libraries,(draft) s- 69).

  Organisering:

- **Nuovo Soggettario**
  http://thes.bncf.firenze.sbn.it/index_eng.html

  Bakgrunn: Italiensk generell tesaurus. Utviklet i samsvar med IFLAs anbefalinger og andre internasjonale standarder for emneindeksering.

Struktur: Kan brukes både post- og pre-koordinert.
Omfang: Ca 60 000

Mapping: Dewey er viktig som bro mellom Nuevo Soggettario og emnesystemer på andre språk. http://eprints.rclis.org/10077/

Organisering: National Central Library of Florence (BNCF) har redaksjonsansvaret. Tesaurusen utvikles I samarbeid med den italienske nasjonalbibliografien, Bibliografia Nazionale Italiana, (BNI) http://www.bncf.firenze.sbn.it/pagina.php?id=198