NorNa – tool som et hybridsystem

av

Øivin Andersen

1. Innledning

Denne artikkelen er en gjennomgang av to grunnleggende aktivitetsområder som ligger til grunn for etableringen av en dokumentasjonsbasert terminologisk database som både skal tjene som et verktøy for terminologisk forskning og utvikling på den ene siden, og et tverrspråklig gjenfinningssystem for fagtekster for de nordiske språkene svensk, dansk, norsk og finsk på den andre siden. De to områdene er den moderne terminologilæren og bibliotekslæren.

En av hovedhensiktene med denne fremstillingen er å vise de store overlappingene mellom områdene, samtidig som det er viktig å vise de sentrale forskjellene mellom dem.

2. Terminologi vs iod (informasjon og dokumentasjon)

Både terminologilæren og bibliotekslæren (som er en sentral del av iod) har en forholdsvis lang tradisjon. De to aktivitetsområdene har i for stor grad levd sine separate liv og utviklet sine redskaper hver for seg uten i tilstrekkelig grad å trekke veksler på hverandre.

La oss ta utgangspunkt i hvordan de to områdene selv definerer sin virksomhet:

Iod-vitenskap: *Vitenskapen om dokumentasjonssystemer og deres informasjonsforutsetninger, betingelser og konsekvenser, i.e. systematisering og organisering av kunnskap for senere gjenfinning* (Kongsbakk 2003).

Terminologilære: *[Læren om] de språklige midlene på leksikalsk plan som gjør et fagområdes begrepsapparat kommunikabelt.* (Laurén et al 1997).

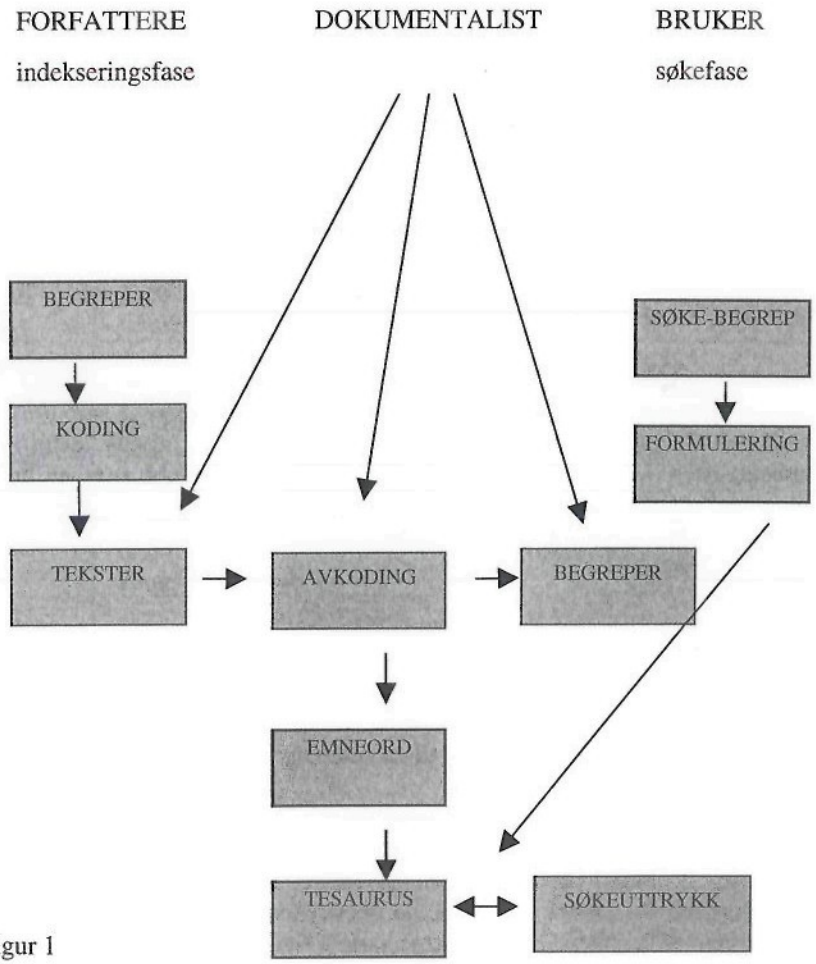
Det sentrale stikkordet i iod-definisjonen er *informasjon*. Iod har lang tradisjon for å være opptatt av de informasjonsprosesserende sidene av leksikalsk strukturering, og dette fokuset er fremdeles sentralt. Når det gjelder terminologilæren, derimot, har begrepet *kommunikasjon* hele tiden vært sentralt. Det mest sentrale her er å gjøre domenespesifikke fagområders subspråk forståelig ikke bare blant fagfolk innen samme domene, men også mellom fagfolk innen ulike domener og (ikke minst) mellom fagfolk og allmennheten.

En annen viktig forskjell er at terminologilæren legger vekt det intime og systematiske samspillet mellom det språklige og det kunnskapsmessige, mens iod legger mest vekt på det gjenfinningsmessige. Tradisjonelt har derfor terminologilæren hatt et nærmere forhold til språkvitenskapen og har vært sett på som en disiplin innen anvendt lingvistikk, til forskjell fra iod. Det som er felles i disse definisjonene er vektleggingen av leksikalsk og begrepsmessig strukturering.

Terminologilæren har likevel ved flere anledninger påpekt likhetene mellom de to områdene (Wüster 1974). Denne fremstillingen bygger på noen av Wüsters observasjoner samtidig som de mer moderne manifestasjonene av dette illustreres ved hjelp av sentrale egenskaper ved en dokumentasjonstesaurus og en terminologisk database. En av de mest sentrale og markante innen iod er Wersig (1985). En del sentrale aspekter vedrørende oppbyggingen av dokumentasjonstesauri er hentet fra ham.

3. Hovedfaser i dokumentasjonsprosessen

For å anskueliggjøre det egenartede utgangspunktet for en dokumentasjonstesaurus er det nødvendig å vise noen av de mest sentrale fasene i selve dokumentasjonsprosessen (modifisert etter Wersig op.cit.:15f):



Figur 1

Oversikten i figur 1 består av tre sett av aktører: forfattere av dokumentene som skal danne grunnlaget for tesaurusen, brukerne, som skal gjenfinne dokumentene i overensstemmelse med sitt informasjonsbehov, og dokumentalistene, som skal formidle mellom forfatterteksteне og brukerne.

Videre kan vi skille mellom indekseringsfasen (venstre side av figur 1) og søkefasen (høyre side) av prosessen.

Både innen allmennspråklige og fagspråklige sammenhenger er som kjent forfattere høyst forskjellige både når det gjelder konseptualisering, formuleringsmåte, ordvalg, inkludert bruk av terminologi, stilistikk og tekststrategiske virkemidler for å kode emner i tekster. Disse forskjellene resulterer vanligvis i et heterogent sett av tekster som ene og alene danner det empiriske grunnlaget for en dokumentasjonsbasert tesaurus. Denne heterogeniteten representerer en stor utfordring for dokumentalisten, som har som oppgave å trekke ut emneord fra disse tekstene og å strukturere dem videre i en hierarkisk ordnet temastruktur, dvs. en dokumentasjonstesaurus.

Når det gjelder indekseringsfasen må dokumentalisten forstå innholdet i tekstene for å kunne avkode dem, dvs. ha kjennskap til de faglige begrepene som brukes. Et mål er å oppnå størst mulig homogenitet. For å oppnå dette må dokumentalisten identifisere ord som er brukt synonymt, ord som er brukt polysemt og homonyme/homografe ord. Dette er grunnlaget for å kunne etablere en vokabularkontroll, noe som er helt nødvendig for et effektivt gjenfinningssystem.

Ved hjelp av en tekstuell innholdsanalyse skal de relevante begrepene munne ut i et sett av emneord, som på en mest mulig effektiv måte skal peke på de dokumentene som de er indeksert fra. Ut fra prinsipper om hvor generelle eller spesifikke emneordene er skal de så struktureres i et sett av hierarkier hvor hierarkiene også skal knyttes sammen til en helhetlig tesaurus.

Når det gjelder søkefasen er også brukerne av et system ofte preget av heterogenitet. De har forskjellig kunnskapsbakgrunn, sosial bakgrunn, kompetanse i bruk av gjenfinningssystemer og ikke minst, ulike informasjonsbehov. Derfor vil også de mer eller mindre klart formulerte søkebegrepene som anvendes av brukerne variere sterkt.

Som figur 1 forsøker å illustrere vil en strukturert tesaurus kunne fungere som en rettesnor for søkeren på det stadiet da hun skal formulere sine søkebegreper. Det er et sentralt mål for dokumentasjonstesauruser at de søkeuttrykkene som brukeren bestemmer seg for å bruke i et dokumentsøk i størst mulig grad stemmer overens med de emneuttrykkene (kalt deskriptorer) som opptrer i dokumentasjonstesaurusen.

4. Forskjeller mellom dokumentasjonstesauruser og terminologidatabaser

I tabell 1 nedenfor har jeg listet noen sentrale og typiske forskjeller og likheter mellom dokumentasjonstesauri og terminologidatabaser.

	DOKUMENTASJONS TESAURI	TERMINOLOGIDATA BASER
a.	angir ufullstendige hierarkier	skal ideelt angi fullstendige hierarkier
b.	er deskriptiv	er ofte normativ
c.	deskriptorer refererer til språklig verden	termer refererer til utenom-språklig verden
d.	deskriptorer refererer til finitt antall forekomster/belegg	termer refererer til potensielt sett infinitt antall referenter
e.	deskriptorer er kontekstsensitive	termer er kontekstfrie
f.	angir emneord/deskriptorer	angir termer
g.	foretas av bibliotekarer	foretas av terminologer og fagfolk
h.	angir spesifikasjonsnivåer	angir spesifikasjonsnivåer
i.	skiller som regel ikke mellom generiske og partitive hierarkier	skiller mellom generiske og partitive hierarkier
j.	angir ikke grad av faglighet	angir grad av faglighet
k.	bygger på relevant dokumentsett	bygger på fagspråkskorpora
l.	er ikke representative	er ideelt representative
m.	kan ikke danne nye termer	kan danne nye termer

En av de viktigste skillene er at hierarkiene i dokumentasjonstesauri er ufullstendige (a.). Den viktigste grunnen til det er at det settet av dokumenter som tesaurusen bygger på som regel ikke inneholder dokumentsett som er representative for noe fagdomene. Ofte vil hele nivåer eller rekker i hierarkiene mangle, og som regel vil mange rekker være ufullstendige i forhold til et terminologisk begrepssystem (jf. Wright et al 1997).

En dokumentasjonstesaurus vil ofte være deskriptiv i den forstand at de emneordene som er indeksert fra dokumentsettet ikke har fått noen andre etiketter enn de som opptrer i dokumentene (jf. b.). Det betyr at dokumentalisten ikke har mulighet til å endre eller "forbedre" emneordene som skal inngå i tesaurusen. Imidlertid vil indekserne utføre en slags intern normativ virksomhet i den prosessen hvor emneordene som skal inngå i den hierarkiske delen av en tesaurus tildeles status som deskriptorer. I terminologisk sammenheng er forholdet ofte ganske annerledes. Et av de mest markante kjennetegn ved den moderne terminologivirksomheten er nettopp muligheten til normering av terminologi for å oppnå enhetlig faglig kommunikasjon. Prinsipper for termdanningsvirksomhet er derfor en meget sentral gren av den moderne terminologilæren.

De utvalgte deskriptorene i en tesaurus er definert ved at de refererer til et finitt sett av forekomster (eller belegg) i det vi kaller en språklig verden, dvs. det bestemte settet av dokumenter som deskriptorene er indeksert fra (jf. punkt c. og d.). Termer derimot refererer til den utenomspråklige verden, og det typiske her er at antallet potensielle referenter er infinitt. Dette medfører at deskriptorene i en tesaurus nødvendigvis er kontekstsensitive (jf. punkt e.) mens det klassiske idealet i terminologilæren er størst mulig grad av kontekstfrihet (selv om dette er et omstridt punkt i den moderne terminologiske forskningen).

En annen viktig forskjell mellom de to systemene er at aktiviteter som emneordsindeksering og deskriptorutvalg typisk utføres av bibliotekarer (punkt g.). I en del tilfeller vil bibliotekaren kunne konsultere en fagekspert, i andre tilfeller ikke. I terminologisk sammenheng, derimot, er samarbeidet mellom terminologer og fagfolk et krav til kvalitetssikring av prosessen.

Dokumentationstesauri og terminologiske begrepssystemer har imidlertid det til felles at man kan lese spesifikasjonsnivåer ut av de hierarkiske strukturene (punkt h.). Dette gjelder likevel bare til en viss grad når det gjelder tesaurushierarkiene. Det er to hovedgrunner til dette: For det første er (som tidligere nevnt under punkt a.) tesaurushierarkiene ufullstendige. For det andre blandes som regel generiske (logiske) og partitive (ontologiske) hierarkirelasjoner sammen uten at forskjellen indikeres gjennom begrepsrelasjonene (jf. punkt i.). I et terminologisk begrepshierarki er dette ikke tillatt. De to begrepssystemene kan riktignok blandes i terminologiske fremstillinger av begrepssystemer, men da må de to relasjonstypene angis eksplisitt.

Som en konsekvens av punkt a. og punkt i. kan man normalt ikke uten videre lese grad av faglighet ut fra tesaurushierarkier (jf. punkt j.). Grad av faglighet er derimot en sentral egenskap ved terminologiske hierarkier, og det inngår som sentralt element i fremstillinger av fagspråk.

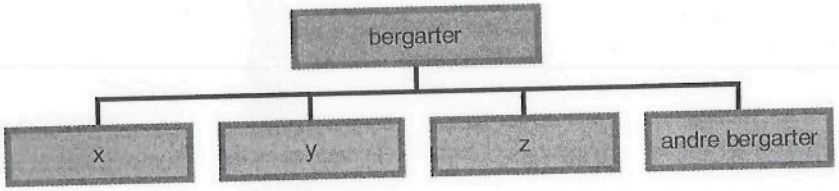
En av de viktigste forskjellene mellom de to systemene er likevel det faktum at dokumentasjonstesauri bygger på et finitt, relevant sett av dokumenter (punkt k.). Mange av disse dokumentene vil være i elektronisk form, men det kan også finnes eldre dokumenter som ikke er det. Nyere dokumentasjon vil imidlertid ofte være kodet i xml-format. Dette vil også i stor grad være tilfelle for fagspråkskorpora, men formålet med et fagspråkskorpus er ganske annerledes enn et dokumentsett for tesauruskonstruksjon.

Den moderne terminologivitenskapen er helt avhengig av gode og representative korpora for å kunne undersøke alle aspekter av terminologibruk i fagtekster. Disse tekstene annoteres på ulike måter avhengig av hvilke aspekter av terminologi (termbruk, betydninger, definisjoner, etc.) som ønskes belyst gjennom korpuset. I et dokumentsett for tesauruskonstruksjon er siktemålet mer endimensjonalt: å finne de emneordene som på en mest effektiv måte peker på sentrale temaområder i dokumentene.

Likevel er det et viktig punkt i NorNa-prosjektet at tesaurusdokumenter er en viktig datakilde for terminologisk korpusbygging. I mange tilfeller representerer de viktige faglige dokumenter som kan inkorporeres i terminologiske tekstkorpora og annoteres videre. Med andre ord søker prosjektet å utnytte den

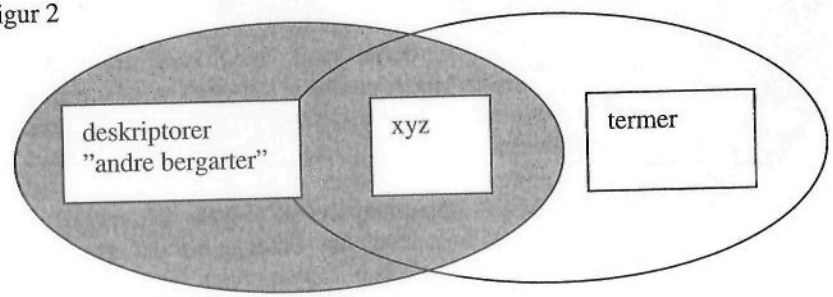
synergieffekten som dette kan gi. Men et tesaurusdokumentsett er ikke representativt for noe fagområde (jf. punkt 1.) og må derfor bearbeides og analyseres videre hvis det skal inngå i et terminologisk korpus for å danne grunnlag for en terminologisk database.

Selv om man ikke kan danne nye termer i en dokumentasjonstesaurus vil de deskriptorene som utgjør tesaurusens systematiske del likevel være en meget god kilde til termidentifikasjon i et terminologisk korpus. De har jo vært igjennom en vokabularkontroll som har mange likhetstrekk med den terminologiske analysen. Hvis man for eksempel tar et tesaurushierarki av bergarter hvor **x**, **y** og **z** både er termer og deskriptorer, mens "andre bergarter" er deskriptor, men ikke term (jf. figur 2 nedenfor), vil forholdet se ut som i venndiagrammet nedenfor (figur 3):



I figur 3 vil x, y og z befinne seg i snittmengden mellom deskriptormengden og termmengden, mens "andre bergarter" er nyttig som deskriptor uten å ha termstatus.

Figur 2



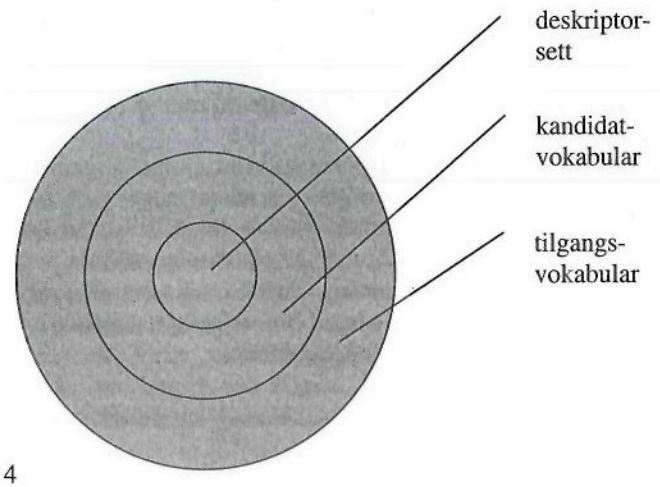
Figur 3

5. Grunnleggende begreper innen dokumentasjonstesauri

I denne delen av fremstillingen vil jeg ta for meg noen av de mest sentrale begrepene i tesaurusbygging.

5.1 Emneord og deskriptorutvelgelse

For å fremstille den delen av dokumentasjonsprosessen som omfatter deskriptorutvelgelse har jeg valgt å illustrere ved hjelp at et venndiagram med konsentriske sirkler. Den ytterste sirkelen angir den totale mengden av emneord som indekseres fra et sett av dokumenter. Dette kalles tilgangsvokabularet. Fra denne helmengden velges de emneordene som er tiltenkt deskriptorstatus i tesaurusen. Denne delmengden kalles kandidatvokabularet. Den innerste sirkelen angir den mengden av deskriptorer som skal emnestruktureres til en dokumentasjonstesaurus. Denne delmengden er ikke noe statisk og "endelig". I praksis vil deskriptorsettet endres avhengig av hvor effektiv tesaurusen i praksis viser seg å være.



Figur 4

Et av de mest sentrale begrepene i dokumentasjonstesaurus er *emne*. Et emne er det settet av enheter som angir det "viktigste" innholdet i et dokument/sett av dokumenter. Emner uttrykkes i *emneord* (eller *indeksord*) og resulterer i en emneordliste. For at denne listen skal kunne bli håndterlig må den underlegges vokabularkontroll. Et *kontrollert vokabular* er en form for semantisk struktur som skal kontrollere synonyme og polyseme ord, skille mellom homonymer og lenke sammen emneord hvis betydning er relatert til hverandre. Et slikt vokabular kalles et dokumentasjonsspråk.

De emneordene som skal brukes i den systematiske delen av en dokumentasjonstesaurus kalles deskriptorer. Settet av deskriptorer er altså en ekte delmengde av settet av emneord. Et emneord som klassifikasjon er polysemt og må derfor disambigueres. Dette kan gjøres ved såkalte betydningsindikatorer (scope notes), som settes i parentes etter ordet: *klassifikasjon* (prosess) og *klassifikasjon* (produkt). Betydningsindikatorer brukes også hvis et og samme emneord forekommer i to eller flere hierarkier: *hest* (turnsport) og *hest* (biologi). Hvis to eller flere emneord er synonyme vil dokumentalisten som regel velge ett av emneordene og utelukke de andre fra den systematiske delen av dokumentasjonstesaurusen. Hvis *språkvitenskap* ansees for å være en mer effektiv søketerm enn synonymet *lingvistikk* vil førstnevnte emneord få status som deskriptor. Synonymet *lingvistikk* vil da kun opptre i den alfabetiske delen av tesaurusen med peker mot *språkvitenskap*. Dette gjøres uavhengig av hvordan det rent terminologiske forholdet mellom de to ordene er.

Utvalget av deskriptorer for dokumentasjonstesauri er underlagt ganske strenge krav. Når dokumentalisten arbeider med deskriptorutvalg er det i hovedsak tre typer grunnlagsmateriale som brukes: Det første er bibliografiske klassifikasjonssystemer, som for eksempel UDK, et system utviklet av Melville Dewey på 1870-tallet. Et hovedproblem med UDK er at systemet er endimensjonalt og monohierarkisk. I en tid hvor faglige aktivitetsområder i økende grad overlapper er dette et voksende problem.

I tillegg til UDK brukes også andre emneordlister og tesauruser fra beslektede dokumentasjonsområder i tesaurusarbeidet.

5.2 Relasjonsbegreper

I 1970 ble det utarbeidet et sett av regler for tesaurusbygging på nordiske språk (Hagen 1970). Publikasjonen er lettfattelig skrevet og er en instruktiv innføring i hovedprinsipper som fremdeles gjelder.

Publikasjonen anbefaler at de engelske relasjonsbegrepene brukes. Disse er USE (use, bruk, bruk i stedet), UF (used for, brukt for), BT (broader term, videre deskriptor, mer allment ord), NT (narrower term, snevrere deskriptor, mer spesifikt ord) og RT (related term, beslektet deskriptor, parallelt ord). USE viser, blant annet, til foretrukket synonym: *sekundært batteri* USE *akkumulator*. Den kan også ha andre funksjoner, blant annet for å henvise til en foretrukket stavemåte eller for å ekspandere en forkortelse: *IFF* use *identifikasjonssystemer*. USE-funksjonen kan også brukes for å uttrykke begreper som terminologisk sett ikke er synonymer, men som kan betraktes som synonymer i indekserings- og gjenfinningssammenheng: *semantemer* USE *semantikk*. En rekke mer spesielle funksjoner er nevnt i Hagen (1970:15ff).

UF har et logisk symmetrisk forhold til USE, dvs. enhver USE-relasjon fremkaller den reverserte UF-relasjonen: *akkumulator* UF *sekundært batteri*.
BT angir en relasjon fra en mer spesifikk deskriptor til en mer allmenn deskriptor. Den står også i logisk symmetrisk forhold til NT: *akustisk impedans* BT *impedans*, *impedans* NT *akustisk impedans*.

RT-relasjonen er kanskje den vanskeligste relasjonen i dokumentasjonstesauri. Den angir en relasjon mellom to (eller flere) tematisk relaterte deskriptorer: *akustisk impedans* RT *båndbredde*. Dette er en ekvivalensrelasjon, så samme relasjonen genereres automatisk andre veien: *båndbredde* RT *akustisk impedans*. Ved RT-relasjoner er det viktig at dokumentalisten er streng ved etableringen. "Tematisk relaterthet" er et meget vagt begrep, og en inflasjon av RT-relasjoner i en tesaurus vil virke mot sin hensikt. Dette er også en av de største effektivitetsutfordringene ved såkalte emnekart, hvor samtlige relasjoner er av denne typen (jf. Kongsbakk 2003).

I NorNa-prosjektet har arbeidsgruppen valgt å avvike en del fra dette standardoppsettet. Det er flere grunner til dette. Den viktigste er at prosjektet er et flerspråklig eksperimentelt forskningsprosjekt hvor det er viktig å prøve ut nye ideer hvor terminologisk og lingvistisk innsikt kan bidra til å endre standard oppfatning av tesaurusstrukturer og deres påståtte effektivitet. For det andre skal den resulterende hybridbasen kunne benyttes i terminologisk arbeide til felles beste for de nordiske språkene som deltar i prosjektet.

5.3. Ontologier og kontrollerte vokabularer

Begrepet ontologi har nærmest fått en inflatorisk bruk både innen iod og i mange varianter av språkteknologi. Det stammer opprinnelig fra en gren av filosofien som har tradisjon tilbake til Aristoteles. I filosofien betyr ontologi læren om tingenes og fenomenenes eksistensformer. Overført til lingvistisk semantikk er en ontologi en modell av de aspekter ved den ytre verden som det er relevant å representere på en entydig måte. Begrepet kan også referere til egenskaper ved selve representasjonene. Innen iod er termen *ontologi* motivert ut fra det faktum at det er kun de aspektene ved en representasjon som kan leses av datamaskiner som "eksisterer". Derfor er det et ufravikelig krav til ontologier at de skal kunne forstås av datamaskiner.

5.4. Recall og precision

Et hvert informasjonsgjenfinningssystem må ha et evalueringsmål for systemets effektivitet. For dette formålet bruker man to kvantitative måleenheter: *recall* (R), grad av gjenkalling og *precision* (P), grad av presisjon i søk (jf. Salton et al 1983:164ff). R måler mengden av relevant informasjon som faktisk gjenfinnes i et søk (dvs. antallet relevante enheter som faktisk gjenfinnes delt på det totale antallet relevante enheter som fins i et dokumentsett).

P måler antallet faktisk relevante gjenfunnede elementer i et søk (dvs. antallet faktisk relevante enheter delt på det totale antallet gjenfunnede enheter).

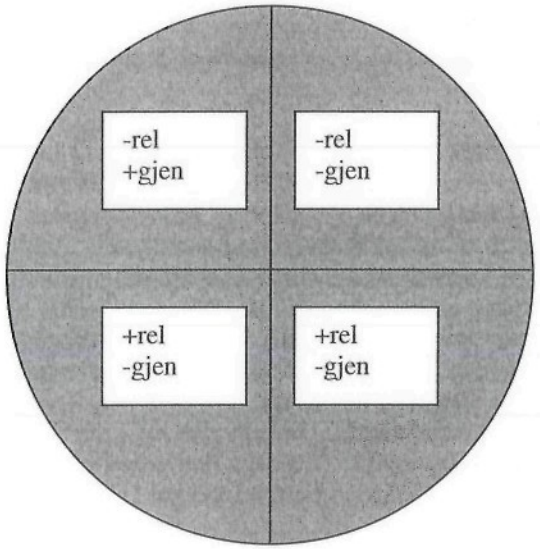
Formelen for R og P er da:

$$R = \frac{+\text{gjenfunnet, +relevant}}{+\text{relevant, totalt}}$$

$$P = \frac{+\text{gjenfunnet, +relevant}}{+\text{gjenfunnet, totalt}}$$

La oss ta utgangspunkt i en dokumentsamling (S) på 200 dokumenter (jf. figur 5 nedenfor):

Recall and precision



Figur 5

Av de 200 dokumentene ble det ved et gitt søk funnet 22 dokumenter. Av de 22 dokumentene var 18 relevante mens 4 var ikke relevante. Det totale antallet relevante dokumenter i søket var anslått til 23. Antallet relevante dokumenter

som ikke ble gjenfunnet utgjør da 5 av de 180 dokumentene som totalt sett ikke ble gjenfunnet:

S = 200
+gjenfunnet, totalt= 22
-gjenfunnet, totalt= 178
+relevant, +gjenfunnet = 18
-relevant. +gjenfunnet = 4
+relevant, totalt = 23
+relevant, -gjenfunnet = (23-18) = 5
-relevant, -gjenfunnet = (180-5) = 175

R = 18 : 23 = 0,782
P = 18 : 22 = 0, 818

6. Konklusjoner

Denne elementære gjennomgangen av noen av grunnbegrepene innen de to disiplinene viser at de to disiplinene iod og terminologi er både like og forskjellige. Hvis vi tar høyde for forskjellene kan vi også utnytte likhetene. Det er nettopp dette NorNa-prosjektet tar sikte på. Det er all grunn til å tro at et databasesystem som tar vare både på effektive søkeord i en fagspråklig tesaurus og på de sentrale termene på de fagdomenene som er representert i en dokumentasjonstesaurus vil ha stor nytteverdi.

7. Referanseliste

Hagen, E. (1970): Norforsks arbeidsgruppe for tesaurusspørsmål: **Regler for bygging av thesauri på nordiske språk**. Nordforsk. Stockholm 1970.

Kongsbakk, I. (2003): **Sømløs kunnskap – Om bruk av emnekart**. Diplomoppgave ved Høgskolen i Oslo, Avdeling for journalistikk, bibliotek –og informasjonsfag. ABM-skrift. Oslo.
http://www.abm-utvikling.no/publisert/ABM-skrift/2004/emnekart.pdf

Laisiepen, K. E. Lutterbeck & K-H Meyer-Uhlenried (1980): **Grundlagen der praktischen Information und Dokumentation**. K.G. Saur. München.

Laurén, Chr., J. Myking & H. Picht (1997): **Terminologi som vetenskapsgren**. Studentlitteratur. Lund.

Salton, Salton, G. & M. J. McGill (1983): **Introduction to Modern Information Retrieval**. McGraw-Hill, Auckland.

Wersig, G. (1985): **Thesaurus-Leitfaden. Eine Einführung in das Thesaurus-Prinzip in Theorie und Praxis**. K.G.Saur. München.

Wright, S.E. & G. Budin (1997): **Handbook of Terminology Management**, vol.1. Basic Aspects of Terminology Management. John Benjamins. Amsterdam.

Wüster, E. (1974): Die Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. I: **Linguistics. An international Review**, nr. 119. Mouton-The Hague, s. 61-107