Redskabsgymnastik med emneord og klassifikation eller frit fald i nettet

Temadag om emnedata på Vejle Bibliotek 27. januar 2005

Refereret af Hanne Sonne Henriksen, Bjarne Christensen, Else Marie Lauridsen Henning Midtgaard Hanssen

Er klassifikation forældet i forhold til emneord? Eller er det tværtimod sådan, at emneord forgår, mens klassifikation består? Eller er begge dele på vej over i historien overflødiggjort af Google og de andre effektive søgemaskiner på nettet?

Hanne Sonne Henriksen
Assistent Det Kgl. Bibliotek
Bjarne Christensen
Afdelingsbibliotekar SDUB
Else Marie Lauridsen
Førstebibliotekar DPB
Henning Midtgaard Hanssen
Førstebibliotekar SB

Det vil være synd at sige, at emneregistrering har været et varmt emne i de senere års biblioteksfaglige debat. Hvis man gennemgår registrene til de seneste 10 årgange af DF Revy, finder man siger og skriver TO artikler under indgangene Emnesøgning og Klassifikation. Med tanke på, hvad der er sket på andre områder i samme periode, vurderede vi derfor sidste år i bestyrelsen for DFs Forum for Registrering, at tiden måtte være inde til at få belyst, hvordan de aktuelle vilkår er for denne traditionelt meget prestigefyldte del af bibliotekernes registreringsvirksomhed: Er klassifikation forældet i forhold til emneord? Eller er det tværtimod sådan, at emneord forgår, mens klassifikation består? Eller er begge dele på vej over i historien overflødiggjort af Google og de andre effektive søgemaskiner på nettet?

Glædeligvis var der så mange, der fandt disse spørgsmål relevante, at det lod sig gøre at samle ca. 70 mennesker til en temadag på Vejle Bibliotek den 27. januar under overskriften "Redskabsgymnastik med emneord og klassifikation eller frit fald i nettet?"

Klassifikation og emneord – dobbeltkonfekt eller supplement? /v. John Kruuse

Den nye formand for Forum for Registrering, Hanne Hørl Hansen, bød velkommen og gav derefter ordet til John Kruuse, for at han kunne fortælle om de erfaringer, han har indhøstet under et projekt, der skulle forbedre emnesøgningsmulighederne i Statsbibliotekets (SB) katalog. Projektet er dog for nylig sat i bero, da det er den nuværende ledelses politik, at der ikke skal anvendes ressourcer på klassifikation og berigelse med emneord. Af samme grund understregede John Kruuse, at hans udtalelser helt stod for egen regning og ikke repræsenterede SBs officielle holdning til værdien af emnedata i katalogposter.

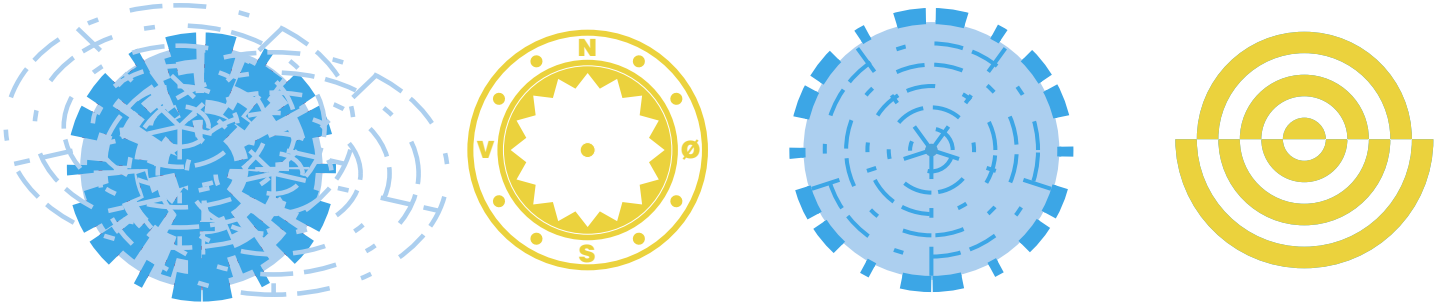
SBs katalog rummer emnedata fra 5 forskellige klassifikationssystemer (UDK, DK5, DDC, NLM og LC samt det lokale FMB-system), 2 verbale emnedatasæt (LCSH og DBCs kontrollerede emneord), plus alle de løse søgeord tilføjet gennem de seneste 25 år. Fra 1998-2002 benyttes dog - forskelligt fra sprog og fag - alene DK5 + DBC-emneord til danske bøger og DDC+LCSH til udenlandske. Siden maj 2004 har man udelukkende genbrugt eksisterende emnedata i DBC- og MLT-poster, så (afhængigt af fag) er mellem 5 og 50 % af alle nyindkøbte udenlandske bøger uden emnedata i katalogen.

Alle emnedata i SBs katalog behandles som autoritetsdata, så et klik i posten tilsyneladende fører til alle andre bøger om samme emne i katalogen. Allerede nu findes 48.000 forskellige DDC-klassemærker, 70.000 forskellige LCSH-emneord, 22.000 forskellige DK5-klassemærker (inkl. alfabetiske underdelinger., tillægstal m.m.) og 16.000 forskellige DBC-emneord, så det blotte antal nødvendiggør mange forskellige søgninger for at ramme selv et meget snævert betragtet fagområde. Et betydeligt antal DDC-klassemærker og LCSH-emneord optræder i flere former (og dermed også i forskellige autoritetsposter), fordi de stammer fra genbrugsposter med varierende formateringspraksis.

Både DDC og DK5 er - for slutbrugeren - meget lukkede klassifikationssystemer. DK5 kan overskues, men man skal have Tante Grøn og Tante Brun ved hånden. DDC har efterhånden udviklet sig til mere at være et regelsæt for, hvordan et klassemærke konstrueres korrekt, end en tabel over mulige klassemærker, og selv om der også her eksisterer et udmærket indeks, tager det ikke hensyn til DDCs mangfoldighed af mulige tillægstal. For mange fag er den FMB- og UDK-klassificerede ældre bogsamling i Statsbiblioteket stadig af stor interesse, men også her mangler brugeren af det virtuelle bibliotek et værktøj til at udpege de rigtige klassemærker før søgning.

DBC-emneord er umiddelbart forståelige, men er som enkeltord i de fleste tilfælde for uspecifikke, LCSH-emneord er meget præcise og let genkendelige. Men for begge systemer savnes en oversigt over hvilke emneord, man kan forvente at træffe i en bestemt faglig sammenhæng. Det meget store antal eksisterende LCSH-emneord - mere end 220.000, hver med mulighed for op imod 100 forskellige subheadings - medfører desværre også, at mange ens bøger LCSH-klassificeres temmelig forskelligt.

Alle disse problemer kan løses på en simpel måde: Vi laver en konkordans mellem klassifikationssystemerne DDC, DK5,



Det overordnede mål med emneregistreringen er, at det skal være et let forståeligt og let anvendeligt værktøj til lånerne. Der skal derfor anvendes ord, som lånerne naturligt ville anvende.

NLM, FMB og UDK (LC er bevidst udeladt) og hægter de eksisterende emneord (LCSH+DBC) op på de fagområder, som konkordansen opdeles i.

Alle emnesøgningsmuligheder i bibliotekskatalogen præsenteres i et browsebart hierarki, hvor vi selv definerer finheden. De yderste grene er maskingenererede websider, hvor alle klassifikationskoder er oversat til klikbar klartekst, og hvor evt. trunkeringer, årstalsafgrænsninger m.m. er indlagt i søgestrengen "bag" den klikbare tekst. Ved at søge på emnedata "udefra" i stedet for at klikke på uforståelige emnekoder i en post opnår vi to ting: at slutbrugeren får de rigtige klassemærker til rådighed for søgning straks, og at de små variationer i delfelt-formateringen af klassemærke eller emneord i posten, der giver forskellige autoritetsposter, søges under ét.

For fagområder, hvor den monografiske litteratur har stor betydning, kan siderne let udvides til også at omfatte søgning i fælleskataloger som fx bibliotek.dk (hvor DDC og LCSH via genbrug har fundet vej til poster fra mange biblioteker, der ikke selv aktivt benytter de to systemer). Men søgestrengene skal naturligvis tilpasses den større variation, som specielt LCSH udviser i meget store bibliotekskataloger.

Konklusionen er derfor: Ingen emnedata er overflødige, tværtimod supplerer mange hinanden i et uventet stort omfang. Imidlertid har slutbrugeren ringe chance for at udnytte dem i den form, de i dag foreligger på i SBs katalog, i bibliotek.dk og andre store bibliotekskataloger. (Se også J.Kruuses uddybende artikel i dette nr.)

Hvordan søger brugerne egentligt? /v. Kirsten Larsen

Kirsten Larsen fra Dansk BiblioteksCenter delagtiggjorde forsamlingen i sine erfaringer med, hvordan forskellige typer af brugere - både den professionelle og slutbrugeren – søger i bibliotek.dk og Netpunkt.

En analyse af en enkelt dags søgninger i bibliotek.dk (ca. 20.500) viser, at der ikke foretages så mange fritekstsøgninger, som man kunne forvente. Faktisk var kun ca. 7 % af samtlige søgninger på fritekst. Derimod var der mange specifikke forfatter- og titelsøgninger: henholdsvis 34 og 20%. Emnesøgninger tegnede sig for 11%. Videresøgninger fra et allerede fundet forfatternavn eller emne udgjorde henholdsvis 5 og 11%. Slutbrugerne søger altså ofte temmelig præcist.

Der er nogle overraskende forskelle på, hvordan der søges i bibliotek.dk og i DanBib via Netpunkt (sidstnævnte anvendes som bekendt hovedsagelig af bibliotekerne). I Netpunkt var 48% af søgningerne på en enkelt dag med ca. 37.600 søgninger fritekstsøgning, og mange søgekoder var slet ikke anvendt. Resultatet viser, at brugergrænsefladens arkitektur har stor indflydelse på, hvordan der bliver søgt.

Kirsten Larsens indlæg har i bearbejdet form kunnet læses i DF Revy 2005:3, s. 6-8, og derfor skal kun nogle hovedpunkter refereres her. Centralt i hendes indlæg var arbejdet med at udvikle det emnehierarki, som Biblioteksstyrelsen har besluttet skal sættes i drift sammen med en ny brugergrænseflade i bibliotek.dk i efteråret 2005.

Emnehierarkiet skal hjælpe søgningen på vej, hvis slutbrugeren ikke har et klart defineret informationsbehov, finde overblikslitteratur, hjælpe med at finde frem til enkelte gode titler om et emne og afhjælpe stavevanskeligheder med "forprogrammerede" søgninger.

Eksisterende folkebiblioteksdata bliver kernen - der tilføjes altså ikke nye data. Søge-URL'erne skal kombinere klassifikation og emneord med titler og lidt fritekst, hvilket vil kunne give ret præcise søgninger. Ordhierarkiet hentes fra mange kilder også uden for den traditionelle biblioteksverden, og der arbejdes med at skabe en universel nutidig sprogbrug.

Emneregistrering i praksis 1: Handelshøjskolens Bibliotek, København /v. Anita Sørensen

Handelshøjskolens Bibliotek (HBK) anvender mange ressourcer på emneregistrering. Der sættes i dag danske emneord på alle poster – dog ikke store pakker af e-bøger som f.eks. Ebrary. Der anvendes både kontrollerede og ukontrollerede emneord. Tidligere anvendtes også UDK-klassifikation, men fra 1997 tilknyttes kun emneord.

HBK har opbygget en emneordsbase med bibliotekets eget hierarkiske emneordssystem med synonymer, overordnede og underordnede emneord, beslægtede

begreber og engelske emneord. Basen rummer i dag mere end 6000 forskellige emneordsposter. Problemet vedrørende emneord der kan anvendes forskelligt inden for forskellige fag, søges løst ved at anvende strenge, hvor ordene så indgår i forskellige sammenhænge.

Det overordnede mål med emneregistreringen er, at det skal være et let forståeligt og let anvendeligt værktøj til lånerne. Der skal derfor anvendes ord, som lånerne naturligt ville anvende. Man forsøger dog samtidig at være stringent i valget af emneord, og disse to mål kan stride lidt mod hinanden. Endeligt skal det være muligt via en tesaurus at vise beslægtede emneord.

Fra sommeren 2004 har man indført en søgeskærm for emnedata, hvor der er to muligheder:
1) Bred emnesøgning, hvor man søger direkte i hele databasen og direkte får vist de bibliografiske poster.
2) Præcis emnesøgning, hvor man søger i emneordsbasen. Der søges kun i de kontrollerede emneord. Ved søgning i denne base fremvises emneordsposten, hvor man også kan se evt. overordnede og underordnede emneord – og videresøge direkte fra disse. Der fremvises ligeledes synonymer og se-også henvisninger til beslægtede emneord.

Fokusgruppeundersøgelser har vist, at emneordssystemet er et godt værktøj, men det har været svært at formidle, hvordan det skal anvendes. Forskellen mellem de to forskellige typer af emnesøgning er ikke indlysende for brugerne. Derfor er det tvivlsomt om man i øjeblikket har opnået sit mål. Man har et godt katalog for de professionelle brugere, men det er nødvendigt i fremtiden at skabe en bedre grænseflade med en mere udviklet tesaurusfunktion og med en udvidet bruger/system-dialog.

Det ville også være ønskeligt med andre faglige indgange, f.eks. pr. semester, fag eller opgave. Det kunne også være en mulighed i fremtiden med en decideret engelsk søgeskærm til emnesøgning, hvor man kunne søge på de engelske termer, som ligger i emneordsposterne.

Anita Sørensens indlæg kan studeres nærmere i DF Revy 2005:3, s. 10-12.

Emneregistrering i praksis 2: Syddansk Universitetsbibliotek /v. Poul Hynding

Syddansk Universitetsbibliotek (SDUB) har altid gjort meget ud af emneregistreringen. Man anvender en tilpasset udgave af SAB-klassifikationen. I dele af systemet er den beriget med UDK og NLM-klassifikation. Helt fra kortkartotekets tid har man både anvendt klassifikation og verbale henvisninger til denne. I slutningen af 1980'erne opdelte man systematikken i undergrupper. Der skete desuden en opdeling af de fleste grupper i flere niveauer med undergrupper. F.eks. blev den klassifikation, som svarede til grammatik, underopdelt i de forskellige aspekter af grammatik med en lang række klassifikationskoder med tilhørende verbale henvisninger. Dette gav mulighed for den findeling, der eksisterer i dag.

I det nuværende biblioteksystem (Horizon) ligger SAB-klassifikationen med tilhørende verbale henvisninger som autoritetsposter. Dette har givet nogle begrænsninger i måden, man kan søge på de verbale henvisninger, men har også gjort diverse opdateringer nemmere.

Den seneste udvikling, som foregik i 2003 og 2004, er en berigelse af selve SAB-klassifikation med en lang række enkeltord, som gør en kombinatorisk søgning mulig – enten via en fritekstsøgning (basissøgning) eller via en speciel søgevej for emneord. Disse enkeltord afspejler i høj grad hierarkiet i emneordssystemet, dvs. at grupper nede i hierarkiet også har emneord fra den/de overordnede grupper. F.eks. er der på den SAB-klassifikation, der dækker komedie, som er en undergruppe til drama, tilføjet både emneord fra komedie og fra overgruppen drama. I forbindelse med denne berigelse har der været nogle overvejelser om, hvordan emneord skal udformes som enkeltord – i modsætning til de verbale henvisninger, der ligger som strenge. Det har også været nødvendigt at overveje, hvor langt man skal gå med at berige med synonymer og nært beslægtede ord. Det vigtigste er at finde de formuleringer, som lånerne vil anvende.

Der er planer om at gøre grænsefladen til emnesøgningen bedre, og der mangler p.t. en rigtig tesaurus-funktion. Emneordsystemet indeholder hierarkiet via sine emneord, men det ville være et stort fremskridt at kunne fremvise dette hierarki på en direkte måde i en synlig tesaurus.

Emneregistrering i praksis 3: Det Kgl. Bibliotek /v. Henrik Laursen.

På Det Kgl. Bibliotek (KB) findes mange ældre kataloger, der dækker mange tidsperioder, og hvor såvel format som systematik er forskellige. Derfor skal man lede mange steder og desuden have kendskab til, hvordan de forskellige emner er systematiseret i de forskellige tidsperioder. Et værktøj, der måske kan løse disse samsøgningsproblemer, hedder Topic Maps, og dets anvendelighed og muligheder undersøges i øjeblikket i et pilotprojekt.

Topic Maps er en ISO-standard, der bruges til at beskrive videnstrukturer og forbinde disse med informationsressourcer. Topic Maps kan skabe nye måder, hvorpå der kan navigeres rundt i store forbundne datamængder. En Topic Maps applikation kan læse et Topic Map (XML-fil) og på en grafisk anskuelig måde fremvise sammenhænge mellem emner og linke til de dermed forbundne ressourcer.

Topic Maps anvender begreberne Topic, Occurrence og Association. Topic svarer i det aktuelle tilfælde til emnet (klassifikationen). Occurrence svarer til søgningen i basen (REX) efter bøger med den tilhørende klassifikation. Association optræder i pilotprojektet i to typer: dels en association "svarer til", dels en association "se også". Det er muligt at sammensætte de maps, der knytter sig til enkelte kataloger, således at der kan foretages samsøgninger.

Henrik Laursen viste, hvordan man ved hjælp af programmet "Omnigator" kunne præsentere en oversigt over de forskellige typer af topics (en klassifikationsoversigt i de enkelte kataloger). Desuden så vi et eksempel på en samsøgning i katalogerne.

Konklusionen er, at Topic Maps faktisk kan skabe sammenhæng i meget forskellige kataloger. Da det er skalerbart, kan det udbygges med nye kataloger. Brugeren kan også selv aktivt tilpasse det og foretage sit eget valg af fag og perioder. Det er desuden muligt at udbygge systemet med en egentlig tesaurus.

Projektet er som sagt et pilotprojekt, og det var ultimo januar endnu ikke afgjort, om systemet kommer til at overgå til almindelig drift.

Søgemaskiner og emnesøgninger på internettet: Google /v. Erik Høy

Erik Høy fra Københavns Kommunes Biblioteker havde givet sit indlæg om søgemaskinerne undertitlen: Hvad gør de, og hvorfor rammer de så godt?

Google har, som Erik Høy slog fast, sat standarden for søgemaskiner i dag. Man taler ligefrem om "googlificering", der er blevet lidt af et skræmmebillede i biblioteksverdenen. Samtidig råder Google nu over et effektivt indskanningssystem, og der er planer om at indskanne bibliotheksmaterialer i større stil på bl.a. New York Public Library.

Søgemaskiners effektivitet måles i dag ikke så meget på deres evne til at fremfinde dokumenterne som til at sortere på fundene på en hensigtsmæssig måde. Som regel er der alt for mange fund, så det er vigtigt, at de bedste figurerer blandt de 50 første, der præsenteres for brugeren.

Google anvender et sorteringssystem kaldet PageRank (efter en af Googles fædre Larry Page). Det er patenteret, og de præcise detaljer er stadig hemmelige. Der opereres med op mod 100 forskellige relevanskriterier (hvor i dokumentet er søgeordene placeret, hvor hyppigt optræder de, er de med versaler eller fed skrift, hvor tæt står de osv.). Google lægger også stor vægt på, hvor mange links der er til andre sider. I praksis har PageRank vist sig at være de fleste andre sorteringsmåder overlegen.

De fleste brugere synes, at Google er god til at finde de emner, de søger på. Det er hurtigt og enkelt at søge. Der skal ikke anvendes en avanceret søgesyntaks eller avancerede søgebilleder. Derimod vil de professionelle brugere, bibliotekarer eller andre, som søger efter videnskabelige dokumenter ofte have svært ved at finde disse, da de som regel er lavt placeret blandt de mange fund. Man kan dog i nogen grad sætte sig ud over dette ved at bruge søgestrenge med domæne-afgrænsninger.

Erik Høy omtalte til sidst kort Google Scholar. Den er interessant for bibliotekerne, bl.a. fordi den indeholder link til Library Search (WorldCat).

Opsamling og afrunding /v. Erik Thorlund Jepsen

Som de fleste vil vide, er Biblioteksstyrelsens bibliografiske konsulent Erik Thorlund Jepsen en travl mand, men ikke desto mindre havde han sagt ja til at komme til Vejle og afrunde temadagen.

Erik Thorlund Jepsen mente, at man kan tale om tre typer af informationsbehov hos brugerne, og at der er forskel på, hvilke typer af data der skal til for at opfylde behovene. For overskuelighedens skyld er det her sammenfattet i dette skema:

Informationsbehov	Databehov
Verifikative	Bibliografiske data samt evt. meget specifikke emnedata (form)
Bevidst emneafgrænsede	Emneord, klassifikationskoder, angivelse af målgruppe og/eller form samt bibliografiske data
Mudret emneafgrænsede	Som ovenfor, men præsentation af muligheder er essentiel (behov for overblik over emne og emnemæssig behandling i system)

Emnehierarkier baseret på gruppering og søgestrenge (dvs. emneord og klassifikationskoder) er velegnede for brugere, der har mudrede emneafgrænsede informationsbehov og måske også svag system- og/eller søgemæssig viden. Emnesøgninger i Google og på nettet i øvrigt vil – efter Erik Thorlund Jepsens erfaringer – kun give brugbare resultater, hvis der er tale om meget bevidst emneafgrænsede og meget specifikke søgninger. Det samme gælder naturligvis verifikative søgninger. Men der er et vigtigt forbehold: Er målet at finde videnskabelig litteratur, er Google langt fra noget hensigtsmæssigt søgeværktøj.

Erik Thorlund Jepsen underbyggede denne påstand ved at referere nogle tal fra det såkaldte Webtapir-projekt. Dette havde til formål at identificere karakteristika til brug for høstning, søgning og rangordning af videnskabelige dokumenter på nettet. Emnerne var udvalgt inden for området plantebiologi: fotosyntese, herbicid-resistens og plantehormoner. En af de mere bemærkelsesværdige konklusioner var, at Googles rangordning ikke prioriterer videnskabelige dokumenter – tværtimod. Godt nok er Google i stand til at finde dokumenterne, men ved et stort antal søgesvar når de normalt ikke op blandt de 830, som vises. Det skyldes, at over halvdelen af de videnskabelige artikler, der publiceres på nettet, ikke har udgående links.

Som afslutning oplyste Erik Thorlund Jepsen, at dette ville være et af de emner, der tages op på Danmarks Biblioteksskoles og Biblioteksstyrelsens konference den 17. marts om fremtidens emnesøgningsmuligheder i bibliotekerne. Denne konference er blevet planlagt uafhængigt af DFs temadag (dog har arrangørerne talt sammen for at undgå datosammenfald), og det giver håb om, at den diskussion, der blev startet i Vejle, ikke dør ud igen lige med det samme.

Forkortelser for klassifikations- og emneordssystemer

DDC Dewey Decimal Classification
DK5 Dansk Decimalklassedeling, 5. udg.
FMB F.M. Bendtsens klassifikationssystem (lokalt system på Statsbiblioteket)
LC Library of Congress Classification
LCSH Library of Congress Subject Headings
NLM: National Library of Medicine Classification
SAB: Sveriges allmänna bibliotekssystem
UDK: Det Universelle Decimalklassifikationssystem