MLT- og DBC-emnedata – en guldgrube for emnesøgning i forskningsbibliotekerne

Af John Kruuse

For mange fag er SBs ældre bogsamling (FMB og UDK-klassificeret) stadig af stor interesse. Men for alle klassifikationssystemerne gælder, at brugeren af det virtuelle bibliotek ikke har noget værktøj til at udpege de rigtige klassemærker - hverken før eller under søgning.

John Kruuse
Skejbytoften 9
8200 Århus N

Statsbiblioteket benytter udelukkende genbrugte emnedata for alle nye bøger. Første skridt blev taget i januar 1999, hvor alle danske bøger (primært pligtafleverng) kun blev emneklassificeret, hvis det var muligt at finde en genbrugspost med emnedata, og siden maj 2004 har alle nyanskaffede udenlandske bøger fulgt samme procedure.

Den ændrede praksis betyder også, at fagreferenten ikke længere ser "egne" nye bøger, inden de går på plads i magasin eller udlånes.

Nu, hvor den første bestyrtelse har lagt sig, må man nok konstatere, at den ændrede praksis har givet et markant løft på emnedatasiden: aldrig har så mange danske bøger fået både dk5 og DBC-kontrollerede emneord og aldrig har de udenlandske bøger fået så detaljerede klassemærker i form af DDC eller NLM og så præcise emneord i form af Library of Congress Subject Headings (LCSH). Men der er selvfølgelig smuttere, især bøger på tysk, fransk og mere eksotiske sprog er tydelig underrepræsenteret i den "nye" biblioteksbase (som ikke længere hedder SOL).

Autoritetsdata uden autoritet

Selvom biblioteksbasen nu kun fødes fra to kilder, har vi stadig omkring en million ældre titler med det gamle systems klassemærker: FMB og UDK, ligesom man før 1999 alene dk5-klassificerede udenlandske bøger af interesse for Overcentralkatalogen. I perioden 1980-2003 har man i et vist omfang også genbrugt MLT-poster, så vi står nu med et - formentligt enestående (emnemæssigt set) - uensartet katalog med 48.000 forskellige DDC, 70.000 forskellige LCSH, 22.000 forskellige dk5 (incl. alfabetiske underdelinger., tillægstal m.m.) og 16.000 forskellige DBC-emneord, de gamle FMB og UDK er der godt 22.000 af (incl. alf. underdelinger), så det blotte antal nødvendiggør mange forskellige søgninger for at ramme selv et meget snævert betragtet fagområde

Alle emnedata i Statsbibliotekets katalog behandles som autoritetsdata, så et klik i posten burde føre til alle andre bøger om samme emne i katalogen. Men et betydeligt antal DDC og LCSH optræder i flere former (med forskellige autoritetsposter) fordi de stammer fra genbrugsposter med varierende grad af underdeling af DDC-strengen med *b-felter og skiftende delfeltkoder i LCSH-strengen. dk5 har andre, men tilsvarende problemer.

Videresøgning fra en posts autoritetsdata fører derfor kun til en delmængde af de eksisterende poster med samme klassifikation.

Ingen værktøjer til rådighed for emnesøgning

Både DDC og dk5 er - for slutbrugeren - meget lukkede klassifikationssystemer. dk5 kan overskues - hvis man har Tante Grøn og Tante Brun ved hånden. DDC har efterhånden udviklet sig til at være et regelsæt for hvordan et klassemærke konstrueres korrekt fremfor en tabel over mulige klassemærker, og selv om der også her eksisterer et udmærket index, tager det ikke hensyn til DDC's mangfoldighed af mulige tillægstal. For mange fag er SBs ældre bogsamling (FMB og UDK-klassificeret) stadig af stor interesse, men for alle klassifikationssystemerne gælder, at brugeren af det virtuelle bibliotek ikke har noget værktøj til at udpege de rigtige klassemærker - hverken før eller under søgning.

Kun NLM er frit tilgængeligt på nettet, men en praktisk brug hæmmes af, at kun et fåtal af skemaernes mange mulige indgange er repræsenteret i SBs katalog, og den udpræget flade struktur i hver medicinsk speciales gren i hierarkiet leder ikke den almindelige bruger op mod de store grupper, hvor bøgerne "er".

DBC-emneord er umiddelbart forståelige, men er som enkeltord i de fleste tilfælde for uspecifikke. LCSH-emneord er meget præcise og let genkendelige, men i begge tilfælde mangler vi 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 i mod 100 forskellige Subheadings, medfører desværre også, at mange ens bøger LCSH-klassificeres temmelig forskelligt.

En konkordans er løsningen

Som man kan forstå er situationen mere end kritisk, men alligevel ligger løsningen på de fleste af vores problemer lige for: vi skal lave en konkordans mellem klassifikationssystemerne, og vi skal emneklassificere emneordene!

Nu vil den biblioteksvidenskabeligt funderede læser nok indvende, at al forskning (og den er omfattende) har vist, at det ikke er muligt at lave konkordanser mellem forskellige klassifikationssystemer. Det er klart, at det enkelte klassemærke fra f.eks. DDC ikke altid kan matche et og kun et dk5 eller FMB, men for universalbiblioteket SB viser det sig, at en acceptabel (om ikke 1:1) match kan opnås med intervaller af de forskellige klassifikationssystemers "numre". For specifikt "danske" forhold kan DDC måske slet ikke bidrage, og omvendt har dk5 ikke meget at gøre godt med indenfor snævre højvidenskabelige fagområder.

Både DDC og dk5 (som jo er rundet af DDC i 1915) har udviklet sig i takt med den eksisterende litteratur og følger ret snævert eksisterende "fag" - som er lette at identificere indholdet af, og det har vist sig, at SBs gamle FMB og UDK - tildelt efter synsvinklen "emne" for næsten alle områder, uden videre kan matches med korte intervaller af DDC/dk5 i et hierarki med 2-3000 grene.

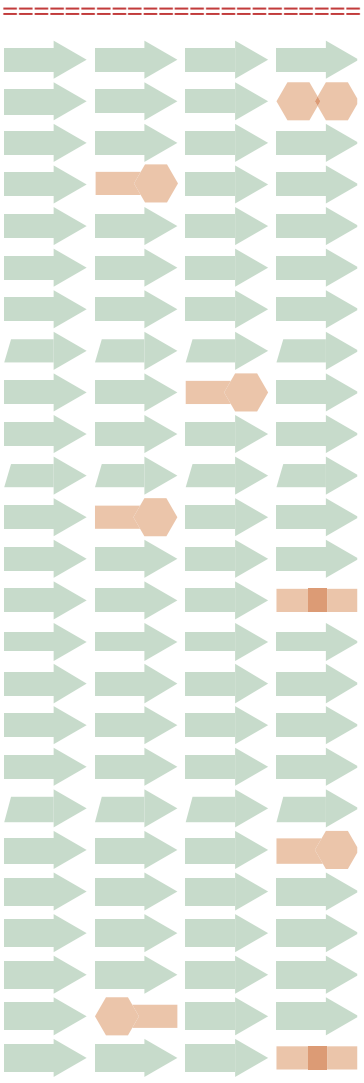
Alle emnesøgningsmuligheder i bibliotekskatalogen præsenteres i et browsebart hierarki, hvor vi altså selv definerer finheden. De yderste grene er maskingenererede websider, hvor alle klassifikationskoder er oversat til klikbar klartekst og hvor evt. trunkeringer, årstalsafgrænsninger m.m. er indlagt i søgestrengen "bag" den klikbare tekst.

Konkordansen konstrueres og vedligeholdes således

Alle poster i bibliotekskatalogen produceret efter 1979 (hvor DDC og LCSH første gang viser sig i MLT-poster) er analyseret for samtidig (parvis) forekomst af DDC, dk5, FMB og LCSH + DBC-emneord. Spændvidden af emnehierarkiets grene er herefter fastlagt efter antallet af repræsenterede DDC, og alle dk5, FMB og LCSH + DBC-emneord, der falder indenfor det spænd af DDC, som den valgte finhed omfatter, optælles maskinelt. Definitionerne på det enkelte kodede emnedatum (DDC, dk5 og FMB) flettes ind i listen (i det omfang den er kendt - betydningen af FMB har vi gemt fra satsen af sidste trykte udgave), så også en manuel korrekturlæsning giver mening. Men det viser sig nu næsten altid, at de klassemærker, der optræder hyppigt sammen, også rent faktisk "falder" sammen rent fag/emnemæssigt. For LCSHs vedkommende er situationen lidt mere uklar - der er meget få hits på den enkelte streng, så man må inkludere de mest karakteristiske ord efter en forventning om flere poster med samme verbale klassifikation i fremtiden.

Alle emnedata er forsynet med en kode, der fortæller om type af klassemærke (af hensyn til søgekoden), om det er realiseret i SBs katalog (hvis ikke: type1) og om det skal søges med eksakt match (type2) eller trunkeret (type3).

En sådan fagopdeling skal foretages een gang for alle. Der kommer hele tiden nye klassemærker, alfabetiske underdeliger af eksisterende klassemærker og nye LCSH og DBC-emneord, men dem kan man håndtere således:



SAMLET LISTE: EMNEDATA CUBAS HISTORIE

XLIS	X972_91	
XTXT	Cuba's historie	
DDC2	[972.91]	History of Cuba
DDC2	[972.9100461]	History of Spaniards
DDC2	[972.9100496]	History of Africans
DDC1	[972.9101]	Early history to 1492
DDC1	[972.9102]	Period of European discovery, exploration, conquest, 1492-1514
DDC1	[972.9103]	1514-1763
DDC1	[972.9104]	1763-1810
DDC1	[972.9105]	1810-1899
DDC2	[972.9106]	1899-
DDC3	[972.9106092*]	Persons involved
DDC1	[972.91061]	Period of American military occupation, 1899-1902
DDC1	[972.91062]	1902-1933
DDC2	[972.91063]	1933-1958
DDC2	[972.91064]	Period of Fidel Castro, 1959-
DDC3	[972.91064092*]	Persons involved
LCS3	[Castro Fidel]	Castro, Fidel
LCS2	[Cuba Politics and government 1959]	Cuba . Politics and government, 1959-
LCS2	[Cuba History Revolution 1959]	Cuba . History, Revolution, 1959
LCS2	[Cuban Missile Crisis 1962]	Cuban Missile Crisis, 1962
LCS2	[History Invasion 1961]	History, Invasion, 1961
dk52	[98.681]	Cuba's historie
dk52	[98.681-99 Castro, Fidel]	Fidel Castro
dk52	[98.681-99 Fernández, Alina]	Aline Fernandez
LCS2	[cubakrisen]	Cubakrisen
FMB2	[Am 49 Cuba]	Cuba - Beskrivelse, samfund, historie
NOTE		

Hver uge laves et udtræk af foregående uges nye poster i katalogen, sorteret efter fagkonto til intern orientering. Hver måned kumuleres ugelisterne, sorteret efter fag (baseret på DDC, dk5 og fagkonto) til orientering for Universitetets biblioteker og forskere og offentliggøres gennem fagets emneguide på SBs hjemmeside. Alle emnedata i de månedlige nyhedslister sammenlignes maskinelt med de enkelte fags "masterliste" og nye elementer flettes automatisk ind i skemaerne. Til slut køres masterlisten gennem et program, der på basis af koderne i venstre spalte genererer XML-udgaven af de skemaer, der indeholder ændringer i forhold til sidste måneds opdatering. Listerne placeres i fagets mappe på et særligt drev og bliver øjeblikkeligt tilgængelige for linket "Faget i bibliotekskatalogen" i den enkelte emneguide på SBs website.

Emnesøgning "udefra" er nødvendig i en Horizon-katalog

Ved at søge på emnedata "udefra" fremfor ved klik 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-formatteringen af klassemærke eller emneord i posten, der giver forskellige autoritetsposter, søges under et. Endvidere løser vi et specielt Horizon-problem: den alm. tilgængelige søgekode for klassifikation (.cl.) er generisk, så den ikke skelner mellem type af klassemærke. Til gengæld søger den ind i klassemærkets tegnfølge, hvor den er brudt af decimalpunktum, *b-underdeling o.lign, og minsandten også på tværs af forskellige klassemærker. En søgning på dk5, f.eks. 66.63.cl. (Farvestoffer + Farvekemi) giver samme resultat som en søgning på 63.66.cl. (Katte) og en DDC-søgning på 196.cl. (Spansk filosofi i nyere tid) resulterer i en stribe poster for børnebøger på finsk, grønlandsk osv. (fra dk5-tillægstal *z 196).

Med sådanne problemer er der ikke noget at sige til, at både slutbrugere (i det omfang de overhovedet kender begrebet emnedata), det vejledende bibliotekspersonale (i det omfang vi overhovedet får lejlighed til at tale med brugerne) og biblioteksledelsen (i et omfang jeg ikke tør nævne) forkaster emnesøgning som værktøj og i stedet accepterer fritekstsøgning. Vi ser også tit brugere benytte Amazon.com til at finde bøger om et givet emne. Man finder også mange relevante titler, som ofte resulterer i fjernlån eller nyindkøb, selvom SB allerede har anskaffet de fleste af fagområdets vigtigste bøger.

Emnesøgning i bibliotek.dk og udenfor

For fagområder, hvor den monografiske litteratur har stor betydning, kan siderne let udvides til også at omfatte søgning i bibliotek.dk (hvor man finder mange DDC+LCSH i MLT-poster, også for alle de andre danske biblioteker, der ikke benytter de to klassifikationssystemer), COPAC o.m.a., om end søgestrengene naturligvis skal tilpasses den større variation, som specielt LCSH udviser i meget store bibliotekskataloger.

En emnesøgeflade til bøger kan præsenteres overfor slutbrugeren på mange måder. Den her udviklede model har indtil nu fundet anvendelse som samlet oversigt over et stort fags litteratur (f.eks. Historie: **www.statsbiblioteket.dk/emneguide/historie/fagisol.xml**) og opsplittet i mange enkeltdiscipliner, hvor tanken i stedet har været at præsentere bogsøgning som en af mange muligheder efter valg af snæver faglig disciplin (f.eks. Kemi: **www.statsbiblioteket.dk/emneguide/kemi/**).
Jo tættere, man kommer fagets tradition for litteratur i bogform, jo mere præcist kan søgefladen skræddersys.

Meget få emnedata er overflødige

Efter snart 20 års arbejde med formidling (indenfor naturvidenskab) er jeg nødt til at gå imod den fremherskende opfattelse af emnedatas manglende nytte. Vores brugere efterspørger stadig - og hyppigt - information, som kun er tilgængelig i bogform, og så hjælper hverken Google eller det overflødighedshorn af e-baser og -tidsskrifter, vi i andre sammenhænge ikke kan klare os uden.

Meget få emnedata er overflødige, mange supplerer hinanden i et uventet stort omfang, men slutbrugeren har ingen chance for at udnytte dem i den form, de foreligger i dag i SBs katalog, i bibliotek.dk og andre store bogkataloger.

Det vil ikke være en uoverkommelig opgave at lave en fælles dansk søgeflade til en væsentlig del af LCs og BLsMLT-emnedata; om vi kan nøjes med DBCs nyudviklede emneflade for danske bøger, må tiden vise. Hvis ikke, har arbejdet med konkordanser allerede vist, at en matchning fag for fag bestemt er indenfor rækkevidde.

Note: Alt arbejde med emnedata har været sat i bero på Statsbiblioteket siden maj 2004. Denne artikel har jeg derfor skrevet som privatperson. Ingen af de fremsatte synspunkter må på nogen måde opfattes som SBs officielle politik på emnedataområdet.

Søgestreng i SBs katalog - som den skal ligge kodet i en webbaseret søgeflade:

I SBs katalog - giver 2 hits:

English language . Foreign elements. Latin

I bibliotek.dk - giver 5 hits:

English language . Foreign elements. Latin

I RedLightGreen - giver 48 hits:

English language . Foreign elements. Latin