# Databrøndens ABC

At opbygge en databrønd er sværere, end man umiddelbart skulle tro. Det er den komplekse sum af rigtig mange små udfordringer, der vokser ud af mange underliggende dataleverandører, hver med deres specielle datastrukturer og -kvalitet. Lidt om brede søgninger og deres forudsætninger.

Af Mogens Sandfær ms@dtic.dtu.dk, Danmarks Tekniske Informationscenter

I dag ønsker de fleste biblioteker at tilbyde deres brugere et enkelt søgefelt, en enkelt søgegrænseflade til hele bibliotekets udbud af informationskilder. Dette kaldes ofte integreret søgning eller mere sigende på engelsk, "single search". Hvad skal der til for at tilbyde dette?


**A. Man skal bruge en søgeportal**, grænsefladen til brugerne med deres web-browsere, mobiler mv. Det er her, det enkelte søgefelt tilbydes, og det samlede søgeresultat fremvises. Som vist på tegningen ovenfor løser søgeportalen sin opgave ved i tur at trække på en eller flere søgemaskiner. Portalens tekniske opgave vokser støt med antallet af søgemaskiner, den skal integrere og harmonisere, især hvis disse arbejder og kommunikerer forskelligt. Under alle omstændigheder har portalen en betydelig designmæssig udfordring. Den skal præsentere et samlet, stort og inhomogent søgeresultat på meget lidt plads. Skal æbler, pærer og pinjekerner præsenteres blandet eller parallelt eller sekventielt? Til eksempel spænder DTU's "single search" over ca. 200 mio. artikler og 200.000 bøger – en faktor 1000 til forskel. Til gengæld er bøgerne måske 20 gange så omfattende som artiklerne – og måske 100 gange så overbliksskabende. Det mest centrale skal udvælges til visning på den altafgørende første svarside, men ofte kan dette betyde, at der kun bliver plads til artikler. Er dette optimalt – også for en førsteårsstuderende? Omkring dette dilemma har der været arbejdet meget – og her er der fortsat meget at udrette. En spændende udfordring uden facitliste.

**B. Man skal bruge en eller flere søgemaskiner** uden egen brugergrænseflade, som søgeportalerne kan udnytte via webservices. Søgemaskinernes datakvalitet og hurtighed er afgørende for brugernes oplevelse af søgeportalen. Trækker man på mange søgemaskiner, uden for egen kontrol, vil man før eller siden løbe ind i problemer med kvalitet og hastighed. Dette blev klart demonstreret af de kommercielle metasøgningssystemer, der blev oversolgt som magiske "single search"-løsninger. At søge på tværs af et stort antal forskellige søgemaskiner virker bare ikke godt nok i praksis. Er svaret så én enkelt søgemaskine? Ja, men det er de færreste beskåret at kunne bygge den. Så løsningen for de fleste bliver at satse på nogle få, der kan tilpasses bibliotekets behov. Kan man selv bestemme, hvad der skal fyldes i søgemaskinen, og hvordan den skal virke, er dette næsten ligeså godt. Måske én for bøger, én for artikler, én for tidsskrifter, én for multimedier el. Bredt dækkende søgemaskiner som de nævnte er afhængige af adgang til et bredt dækkende datasæt af høj kvalitet, altså såkaldte databrønde.

**C. Man skal bruge en eller flere databrønde**, der tilbyder kvalitetsdata til søgemaskinerne. "Databrønd" blev det danske navn for al den databehandling, der skal til, inden et bredt og konsolideret datasæt kan fodres til søgemaskinerne. Bag databrøndene ligger specialiserede dataleverandører i stort tal og med stor variation i det datamateriale, de kan stille til rådighed. Databrøndens opgave er således at indsamle, komplet-

tere, oprense, harmonisere etc. store datamængder fra mange leverandører, der ikke tænkte på "single search" og sammenhæng på tværs, da de designede deres dataformater og -rutiner. En anden hovedopgave er at finde og krydsreferere dubletter, eller at danne værkklynger, som det også kaldes. I mange søgescenarier ønsker man ikke at vise dubletter – i andre er fremfinding og sammenligning af dubletter en væsentlig pointe. Databrønde bør altså ikke foruddiskontere, hvordan deres data vil blive udnyttet af de søgemaskiner og -portaler, der trækker på dem. Bag enhver søgemaskine ligger der således en databrønd, hvad enten den er synlig for omverdenen eller ej. Når der tales om "virtuelle databrønde", tales der faktisk om søgemaskiner, og bag disse ligger der helt reelle databrønde. Mennesker og maskiner, lokaler og lønninger osv.

Man skal altså bruge både A, B og C, men man behøver ikke at bygge og drive det hele selv. Det tunge arbejde i baggrunden kan man deles om. Databrønde laver skjult og neutralt arbejde i baggrunden, men de er afgørende for, hvilke søgemaskiner der kan udbydes, og dermed for, hvilken kvalitet, der kan opnås omkring det eftertragtede ene søgefelt.

Databrønde er opstået i biblioteksregi som et typisk eksempel på samarbejde om fælles "back-office" opgaver. For nylig er der også kommet eksempler på databrønde i kommercielt regi. Men disse databrønde er lukkede og tjener som eksklusivt grundlag for firmaets egen søgemaskine, hvor forretningsmodellen er at sælge så mange billetter som muligt til en og samme søgemaskine.

At opbygge en databrønd er sværere, end man umiddelbart skulle tro. Det er den komplekse sum af rigtig mange små udfordringer, der vokser ud af mange underliggende dataleverandører, hver med deres specielle datastrukturer og deres datakvalitet – og til tider mangel på samme. Udenfor bogområdet følger dataleverandørerne stort set ingen standarder. For tidsskriftforlag er katalogdata som sådan ikke produktet, det er tidsskrifterne, de lever af, og katalogdata er primært produceret til egne portaler og kataloger. Heraf følger, at det ofte er sparsomt med dokumentation, og at den ikke altid overholdes. Tidskrævende "trial and error" spiller en stor rolle, og processen gentager sig, når forlaget beslutter sig for at ændre praksis.

**Danmark i front med åbne brønde**

Det unikke ved den danske situation er, at vi faktisk har to store brede databrønde målrettet bibliotekerne, deres søgebehov og -præferencer. Disse databrønde er velkonsoliderede, af høj kvalitet og har bevist dette gennem mange års stabil drift. En situation, der ofte påkalder sig international opmærksomhed.

Den første er DanBib, der siden 1980erne har været den samlende danske databrønd for metadata om bøger. I DanBib indgår alle danske bibliotekskataloger, den danske nationalbibliografis bogfortegnelse og som gigantisk supplement, bibliotekskatalogen fra Library of Congress. Hermed rummer DanBib næsten alt, man kunne ønske sig at fodre en "library single search" søgemaskine med. For mange vil det snarere være spørgsmålet om at vælge noget fra, inden det sendes til indeksering. DanBib produceres af DBC for Styrelsen for Bibliotek og Medier.

Den anden er DADS, Digital Article Database Service, der siden 1990erne har været den samlende danske databrønd for metadata om artikler i forskningsbibliotekernes tidsskrifter og seriepublikationer. I DADS indgår metadata fra de store globale tidsskriftforlag som Elsevier, Springer mv., der tilsammen står for ca. 80-90 % af artiklerne. Resten, de 10-20 %, udgøres af tidsskriftverdenens "long tail", den lange hale af små forlag, der hver blot udgiver et enkelt eller nogle få tidsskrifter.

Det bliver deres artikler naturligvis ikke mindre vigtige af. DADS omfatter derfor et bredt dækkende datasæt fra tidsskriftagenten Swets, der supplerer de store forlags leverancer. Desuden omfatter DADS brede bibliografiske databaser som Web of Science og Scopus foruden fagspecifikke baser som PubMed, Biosis, Inspec etc. Databrønden DADS rummer således det meste af det, man kunne ønske sig at fodre en "library single search" søgemaskine med . For de fleste vil det utvivlsomt være spørgsmålet om at vælge noget fra, inden det sendes til indeksering. DADS produceres af DTIC for de biblioteker, der benytter den, og hvis behov styrer dens udvikling.

For DanBIb som DADS gælder det endvidere, at producenterne er villige til at dele brøndene med bibliotekerne – f.eks. i form af skræddersyede søgemaskiner, hvorved bibliotekerne kan præge deres foretrukne "single search" løsninger. Med bøger og artikler håndteret af de to danske databrønde, bliver det faktisk ganske overkommeligt at supplere og designe sin egen "single search" løsning.

Allerede i 1999 etablerede DEFF en sådan "single search" løsning for at demonstrere projektets kapacitet. Man koblede tre databrønde, hver med egen søgemaskine, nemlig DanBib, DADS og en særlig brønd for medicinske websider. Det var en kraftfuld demonstration, men forud for sin tid. Den blev lukket ned efter at have løst sin opgave: at vise hvad DEFF's vision gik ud på. Nu er tiden blevet moden – og heldigvis er de to danske databrønde der stadig, og de er fortsat klar til lette bibliotekernes arbejde.