BOK OG BIBLIOTEK

Den usynlige veven

500 milliarder vevsider er utilgjengelige via søkemaskiner

Av Even Flood, *førstebibliotekar*

Den usynlige del av Internett, også kalt Invisible web eller Deep web ble kjent via avisoppslag sommeren 2001 etter en rapport fra firmaet BrightPlanet som ble kilden til mange diskusjoner og avisoverskrifter. Den konkluderte med at 500 milliarder vevsider ikke var tilgjengelig via søkemaskinene, den var "skjult". De presenterte sin egen maskin som gikk i dybden og fisket etter all informasjon på nettet. Begrepet ble også kjent gjennom en meget god bok som kom i 2001: The Invisible Web: Uncovering Information Sources Search Engines Can't See av Gary Price og Chris Sherman, to av de store guruene i denne bransjen.

Rapporten til BrightPlanet er interessant, men ikke særlig god, begrepene for hva som utgjør informasjon er for upresise og den søkemaskinen de selv presenterte holdt definitivt ikke hva den lovet. Men den satte fingeren på et meget viktig punkt: Det er mye informasjon der ute som ikke fanges opp av søkemaskinene. Hvor mye det er kan diskuteres. De siste årene har det vært overraskende stille om temaet. Lite er skrevet og mange websider blir ikke oppdatert. På de seneste "Internett Librarian" konferansene og andre konferanser har det ikke vært et eget tema i det hele tatt. Kanskje fordi det er vanskelig å snakke om noe som er usynlig, det ligger i sakens natur? Allikevel: Den usynlige informasjonen i veven er der. Og det er mye av den! Men først må det bli klarere definert hva det egentlig er snakk om. Og den enkleste definisjonen er den beste:
Det er den informasjonen på web som ikke fanges opp av noen søkemaskin ved direkte søk.
Det kan best illustreres med noen eksempler:
1) Bibliotekkataloger, som Bibliofil og Bibsys. De er små vevsted i det synlige web, men katalogen til Bibliofil har 8 millioner titler og oversikt over 14 millioner eksemplarer som er ikke tilgjengelig fra noen søkemaskin. Lignende tjenester er det over alt.
2) Leksika og oppslagsverk: Encyclopædia Britannica med alle sine artikler, Kunnskapsforlagets leksikon og Caplex er noen.
3) Databaser som ERIC, www.eric.ed.gov/ med over en million henvisninger til artikler og rapporter i pedagogikk. Her er startsiden er synlig, innholdet er ikke det. Pubmed, www.ncbi.nlm.nih.gov/entrez/query.fcgi har hele Medline gratis. Som for Eric: Startsiden er synlig, men alt innholdet (over 6 mill. bibliografiske henvisninger til medisinske artikler) er ikke det.

Ikke dypt nok

Grunnen til at robotene som er grunnlaget til søkemaskinen ikke finner informasjonen kan være mange. For eksempel: En del maskiner går ikke dypt nok in i de enkelte vevstedene. Også: Sider det ikke er noen lenke til blir aldri funnet. Men i denne sammenhengen er det tre andre grunner som er de viktigste til at søkemaskinen ikke har informasjonen.

1) Informasjonen har en ”dynamisk” adresse som blir laget i selve søket, ingen faste adresser søkemaskinen kan plukke opp. Eksempler: Databaser som Bibliofil og Bibsys og Pubmed, for å ta de mest kjente.
2) Artikler har en URL, men de som eier databasen har stanset robotene fra å indeksere siden. De vil selv ha retten til å presentere dataene sine som de selv ønsker. Det finnes standarder for å merke sidene med at de ikke skal legges inn i en søkemaskin. Dette gjelder mange oppslagsverker.
3) Kommersielle informasjonstjenester og betalingstjenester belagt med passord eller andre former for adgangskontroll. Her finner vi klassikerene: Dialog, ISI, Orbit-Questel, STN og realtivet nykommere som Factiva, Encyclopaedia Britannica og OCLC Firstsearch. For ikke å snakke om alle elektroniske tidsskrifter fra de store forlagene, som Science-Direct (Elsevier)

Noen norske eksempler er Norsk Legemiddelhåndbok, www.legemiddelhandboka.no, et omfattende oppslagsverk fra Norsk Lægeforening og andre. Bare få sider øverst i hierarkiet er tilgjengelig fra søkemotorene. Den katolske katekisme er i full tekst på norsk og søkbar på www.katolsk.no/kkk/ men teksten er ikke på søkemaskinene. Eller for å ta et meget lokalt eksempel, Trondheimsbildene i www.trondheimsbilder.no/ er heller ikke med.

Et lite knep for å finne om innholdet i et vevsted du har adressen til er med i den synlige eller den usynlige veven: Gå til Google eller Yahoo!, søk på et relevant ord og avgrens søkingen til vevstedet du vil undersøke. Antall svar (eller mangel på disse) vil si mye om hvor dypt stedet er indeksert.

Oppsummering

Så for å oppsummere litt, hva er i det ”usynlige web”? Databaser, oppslagsverker, håndbøker, leksika, elektroniske bøker, kommersielle betalingstjenester. Og det gir en interessant konklusjon:
Det er her den beste informasjonen er! Materiale som ligger i søkbare oppslagsverker og databaser er verdifullt og kvalitetssikret materiale. Den er evaluert av redaktører, satt inn i en overordnet ramme og sammenheng. Ofte er det ressurssterke organisasjoner som står bak. Og mye av det er fortsatt gratis.

Hvordan finne frem i den usynlige web? Vår jobb som bibliotekarer er å hjelpe folk å finne frem i denne labyrinten. Det er å ha en oversikt over ressursene og vise brukerne hvor de er og hvordan de kan benytte dem.

Søkemaskinene kan faktiske hjelpe, mye av det usynlige er blitt delvis tigjengelig via nye tjenester de siste årene.. Google Scholar, Google Book Search (som er det nye navnet til Google Print) og Amazons A9 lar oss søke inne skjult materiale (bøker, tidsskriftartikler) så innholdet er blitt søkbart og det vi søker på kan skrives ut i kontekst. Selv om hele innholdet ikke er tilgjengelig vet vi at det er der og at det kan bestilles eller kjøpe. Så snart dette er blitt klart kan vi gå videre til Bibsys eller andre kilder og finne materialet der. Google og noen andre søkemotorer tar også i en del tilfelle og søker på søkeordene i relevante og svarene blir en del av svarsettet.

Hva med de kommersielle, passordbeskyttede og dyre (ofte meget dyre) informasjonstjenestene?
Gå til bibliotek som har adgang til tjenestene. Det er meget viktig med nettverk, og her mener jeg personlige nettverk, mellom bibliotekarer fra alle sektorer. Veldig mange fagbibliotek og forskningsbibliotek har disse basene og tidsskriftene og kan et stykke på vei hjelpe med informasjon om dem. Her må vi hjelpe hverandre.

En liten oversikt over gode steder på veven hvor det er mer informasjon:

Det er laget en rekke oversikter og portaler over databaser og lignende som er på den skjulte web og som er gratis tilgjengelige. De er gjerne emneinndelt, men kan også søkes direkte. Her er en liten oversikt over gode vevsteder. Det er også mulig å finne mye informasjon ved å søke på ordene ”deep web” eller ”hidden web” i søkemaskinene, både Google, Yahoo og Clusty gir gode resultater..

Boken til Gary Price og Chris Sherman: ”The Invisible Web” (Medford, N.J. : Information Today, 2001. ISBN: 0-910965-51-x) må nevnes. Det ble laget en meget god vevside for boken: www.invisible-web.net/. Men dessverre, den blir ikke oppdatert og informasjonen både i boken og på vevstedet er dermed flere år gammel. Gary Prices egen oversikt på Freepint: Direct Search: www.freepint.com/gary/direct.htm er like interessant.

Wikipedia har (pr 24/11/05) en bra artikkel om ”Deep Web, med mange lenker: en.wikipedia.org/wiki/Invisible_web

Librarians' Index to the Internet, lii.org/ er en liten (ca 10000), men meget godt indeksert katalog over omfattende vevkataloger og databaser på web, fordelt på 4500 emner. Basen kan søkes via meny, ved å bla i emneområdene alfabetisk og ved å søke direkte i fritekst i beskrivelsen av vevstedene. Den blir regnet som en god inngang til den usynlige veven.

Rapporten som startet det hele: brightplanet.com/technology/deepweb.asp, sammen med søkemaskinen de lanserte samtidig: aip.completeplanet.com/ er begge fremdeles online.

Det er laget en Blog om Deep Web: deepwebresearch.blogspot.com/

Open Directory project skuffet her, de har ikke noen relevant innførsel. Men Yahoo katalogen gir en liten oversikt på dir.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Deep_Web/

Andre metoder for å finne frem? Selv om innholdet er skjult, kan man si mye ut fra svaret på søkinger. Og hvis man stiller dem en masse spørsmål og systematiserer svarene så kan resultatet bli en database som sier en god del. Det er ideen bak prosjektet Qprober: Classification of Hidden-Web Databases through Query Probing, qprober.cs.columbia.edu/ og www1.cs.columbia.edu/~gravano/Papers/2003/tois03.pdf Det er foreløpig bare på forsøksstadiet, men kan bli interessant.

Andre artikler av Even Flood

 Sist oppdatert: 26.01.06