Google Books – Et udkast til et nyt bibliotek

Af Morten Thing 3. november 2009 • I kategorien Teknologi •

http://books.google.dk

Den amerikanske bog- og kulturhistoriker Robert Darnton skrev i en artikel i 2008, at skriftkulturen har ændret sig med en hastighed, som kan tage vejret fra én: 4.300 år fra skrift til bog, 1.150 år fra bog til løse typer, 524 år fra løse typer til internettet, 19 år fra internettet til søgemaskiner, 7 år fra søgemaskiner til Googles algoritmisk styrede relevans-søgning. Han kunne have tilføjet at den sidste udvikling medførte, at nye ord dukkede op i mange sprog: at google, at googlificere. (1)

Selvom Google som søgemaskine i nogle henseender vælger søgemuligheder fra (2) er den på grund af sin søgepower og på grund af sine relevans-algoritmer blevet nummer et søgemaskine i verden. Google har udvidet sin søgeflade (3) med bl.a. Google Scholar, som søger i bøger og artikler af forskere, Google World med en fantastisk kortmaskine og Google Books, som i sin tid blev lanceret som et alternativ til eller måske lige frem døden for den kendte bibliotekstype.

Google Books startede sin tilværelse som et hemmeligt projekt i 2002, men blev offentliggjort på Frankfurt Bogmesse i oktober 2004 under navnet Google Print. (4) Selvom nogle forlag var begyndt at lave e-bøger, var der på dette tidspunkt ikke megen bevægelse i dette marked. Det var faktisk svært at sælge e-bøger. Man diskuterede om det skyldtes, at læserne ikke var indstillet på at læse en hel bog ved skærmen, om det var kontrasten på skærmen der var for stor, eller om det var selve bogens materialitet, der gjorde e-bogs-markedet så trægt. Allerede i december 2004 annoncerede Google et helt nyt format på denne diskussion: Man havde indgået en aftale med universitetsbibliotekerne i Michigan, på Harvard, Stanford og i Oxford (The Bodleian) samt New York Public Library om at digitalisere omkring 15 millioner bind indenfor de næste ti år.

Næsten umiddelbart startede en strid med rettighedsindehaverne (forfatterne og forlæggerne) om mulighederne for at digitalisere og fremvise bøger, som stadig var under copyright-beskyttelse. Det er lidt forskelligt fra land til land, men ofte er en bog beskyttet i hele en forfatters liv plus 70-80 år efter forfatterens død. Det betød jo, at Google måtte nøjes med at digitalisere bøger, som i 2004 var trykt før fx 1854. Det ville da også være en god ting, men næppe noget som tiltalte millioner af læsere verden over.

I 2005 anlagde Authors Guild, the Association of American Publishers og forskellige forfattere og forlæggere sag mod Google. I 2006 fulgte et sagsanlæg mod Google i Frankrig. I oktober 2008 indgik parterne i USA et forlig. Ifølge forliget betaler Google 125 mio $ til rettighedsindehaverne, til advokaterne i sagen og til at opbygge et Book Rights Registry. Man venter på, at retten skal godkende forliget.

Som en del af forliget har Google lavet et website, hvor forfattere og andre rettighedsindehavere af udsolgte, men copyrightbeskyttede bøger, kan indgive krav inden 5.6.2010. De vil få 60 $ pr. bog eller fra 5-15 $ for en del af et værk. Til gengæld får Google lov til at søge i bogens tekst og fremvise korte citater (såkaldte 'snippets') i søgeresultater samt op til 20 % af teksten i en 'preview'. Google må annoncere på disse sider og sælge digitale versioner af bøgerne. Copyrightindehavere vil få 63 % indtægterne fra reklame forbundet med deres værker. Der er mange rettighedsindehavere, som ikke er med i forliget, ligesom udenlandske ikke er blevet spurgt, men i princippet har samme muligheder for at registrere sig som de amerikanske.

I august 2009 dannede en gruppe non-profit-organisationer The Open Book Alliance for at gå imod Google. Amazon, Microsoft og Yahoo overvejede at gå med i alliancen, og den 2. september meldte Amazon sig ind.

Samme dag meddelte Tyskland officielt, at man er modstandere af aftalen mellem Google og forfatterne/forlæggerne. Tysklands justitsminister talte som repræsentant for den tredje største bogproduktion i verden og indvendte, at aftalen overskrider de meget snævrere tyske regler og ikke tager hensyn til tyske interesser. Den 18. september sagde Justitsministeriet, at den foreslåede aftale ikke ville blive godkendt uden ændringer. Den overskred amerikansk anti-trust-lovgivning ved at give Google "de facto eksklusiv ret til den digitale distribution af forældreløse værker". Forældreløse værker er bøger, hvis forfattere er ukendte eller ikke til at finde. En ny aftale skal være i orden inden den 9. november 2009.

Der er således tre forskellige typer bøger i Google Books i dag:

1. Bøger som stadig sælges; hvis forlæggeren vil det, kan de købes som e-bøger gennem Google Books;
2. Bøger som er udsolgt, men som stadig er rettighedsbeskyttede; med mindre rettighedsindehaveren opgiver bogen, kan Google kun vise små citater af det eller de ord, der er søgt på; og endelig
3. Bøger uden copyright, som bliver vist fuldt ud.

I 2006 integrerede Google OCLCs World Cat (en katalog over bøger i hele verden) i søgningen, således at man kan få at vide, hvilket bibliotek i nærheden af én, der har bogen. I Danmark får man som regel en henvisning til bibliotek.dk, hvoraf fremgår hvilke danske biblioteker, der har bogen. Det gælder både bøger, som er scannet og bøger som ikke er scannet. Hertil føjede sig efterhånden en del andre informationer, som har forbedret søgningens

Seneste sprognyheder

4/5	Sprogforskerne fandt en skat i skoven | politiken.dk
27/4	Lille indsats styrker små børns sprog | www.dr.dk
26/4	Lad os komme det danske 'jantekomma' til livs | politiken.dk
23/4	Ud med sproget - Berlingske Mener | www.b.dk
20/4	Unikt runefund i centrum af Odense | videnskab.dk
4/1	John Holm, Pioneer in Linguistics, Dies at 72 | www.nytimes.com
10/8	Young women, give up the vocal fry and reclaim your strong female voice | www.theguardian.com
20/5	Bill Funding Native Language Programs Passes | mtpr.org
17/5	Sounds Of The Pilbara II: Songs In Language finishes recording - WAM - West Australian Music | wam.org.au
13/5	Seven US Senators Introduce Bill to Promote Preservation of American Indian Languages - Native News Online | nativenewsonline.net

FLERE NYHEDER >>>

Verdens sprog på Sprogmuseet.dk på et større kort

Mere i kategorien 'Teknologi'

Zipf
SMS – en ny vår for afrikanske språk?
En e-bog er en ny slags bog!
Maskinoversættelse – en sammenligning af to forskellige metoder

Nye kommentarer

Arturo til Hebraisk: Et genoplivet sprog eller et nyt sprog?
Yunus til Two Turkish Loanwords in Swedish
Sonstige til Den Danske Ordbog på nettet
Mads Haupt til Er det forkert at splitte sammensatte ord?
Monica Scheuer til Jødiske efternavne
jane til Jødiske efternavne
InglêS til Sprogene i Mozambique
Birgit Eggert til Hvad er der blevet af Maren?
Grethe Movsing til Hvad er der blevet af Maren?
Henrik Klindt-Jensen til Ded borrijnholmska måled

Artikler om

muligheder. Der er dog stadig en del problemer med fx fremvisningen af navne, idet det er maskinelt afgjort, hvilket navn som vises. Der kan også være problemer med titler. (5)

I oktober 2009 annoncerede Google, at man havde scannet mere end 10 millioner bøger. Bøgerne scannes ind som de ser ud og fremvises i deres oprindelige typografi. Men de er samtidig OCR-scannet, hvorved det enkelte tegn genkendes, og teksten laves om til en tekstfil, dvs. det er muligt at søge i teksten. Der bruges en teknik, hvor man kan scanne mere end 1.000 sider i timen. Der er kommet mange universitetsbiblioteker til, især amerikanske, men også europæiske som Ghent, Madrid, Lyon, Lausanne, Keio Universitetsbibliotek i Japan og Bayerische Staatsbibliothek. De ser en klar interesse i at kunne tilbyde deres egne studerende den service at kunne få elektronisk adgang til egne bøger. Fra europæisk side, ikke mindst fra fransk side, har der været kritik af en sprogmæssig massiv overvægt af engelske bøger.

I 2006 annoncerede Microsoft, at man ville lancere et lignende projekt kaldet Live Search Books. Projektet blev opgivet i 2008, og den ene million scannede bøger findes nu i Internet Archive, det næststørste arkiv efter Google. I Europa har arkivet Europeana også scannet bøger, men tillige mange andre medier og rummer ca. 3 mio forskellige titler af alle medier. Bibliothéque Nationale i Frankrig har lavet sin egen base, Gallica, som rummer 800.000 bøger og andre medier og øges med 5.000 om måneden. Der er Projekt Gutenberg i Tyskland og Projekt Runeberg i Norden, som har gjorte mange public domain bøger frit tilgængelige på nettet. I USA har The National Yiddish Book Center digitaliseret en større mængde jiddishe bøger, som nu ligger frit tilgængelige i et nyt bladre-software.

Samtidig med denne udvikling har enkelte forlag også markedsført hele bogdatabaser. Først kom Early English Books On-line. Basen indeholder ca. 100.000 bøger fra den første trykte engelske bog i 1475 til ca. 1700. Men planen er at inkludere alle 125.000 bind i de engelske kataloger. Der scannes her fra mikrofilm. Et andet projekt er Ebrary, som indeholder 16.000 bøger trykt efter 1999, som alle er læsbare og søgbare på skærm, men som ikke kan printes i fuld længde. Hertil skal lægges en voksende underskov af digitaliseringsprojekter indenfor de enkelte sprogområder. De fleste nationalbiblioteker i Europa har forskellige projekter, som på længere sigt vil være af stor betydning. Dertil kommer, at mange forlag er begyndt at udbyde flere e-bogstitler. På de større biblioteker kan man i dag typisk få on-line-adgang til mange opslagsværker.

E-bogsudbuddet har bl.a. ventet på en fast trend indenfor hvilket elektronisk format, der skal anvendes. I øjeblikket markedsfører Amazon sin e-bogslæser, Kindle, i Europa. Kindlen understøtter kun Amazons e-bogsformat, så man bliver bundet til at handle e-bøger hos Amazon. Tilsvarende har andre producenter forsøgt på denne måde at få sin del af markedet, helt parallelt til Apples sukces med Itunes og mp3-formatet til musik. Mange biblioteker er også begyndt at låne elektroniske filer ud. Også her findes flere modeller. Enten får man et download, som er frit eller et download, som ødelægger sig selv efter en vis 'lånetid'. Samtidig er e-bogsmarkedet virkelig på vej opad. Det amerikanske salg af e-bøger voksede i første halvår af 2009 fra $ 25,8 mio til $ 37,6 mio. (6)

Det er set fra brugernes synsvinkel et enormt skred gennem Google eller andre baser at få adgang til millionvis af bøger, som bare er et klik eller to væk. Man kan for bøger, som er uden copyright, downloade en gratis PDF-kopi. Man kan således opbygge sit eget e-bibliotek helt gratis.

Problemerne ved Google Books er i nogle henseender de samme, som dukker op i andre digitaliseringsprojekter, hvor scanning foretages automatisk og ikke gennemgås af et kritisk menneskeblik. Søger man fx på: <Oehlenschläger Poetiske Skrifter>, får man følgende fund:

> **Poetiske skrifter, Bind 25-26** Af Adam Gottlob Oehlenschläger,

men hvad man i virkeligheden får er bd. 31 fra 1862 af hans Poetiske Skrifter og ikke bd. 25-26. Scanningen er fin nok, i billeder af de gotiske bogstaver. Men maskinens automatiske OCR-scanning af overskrifter er guld værd: fx

> 'rane fortcetler e вода i Äongäljallen' for 'Hrane fortæller en saga i Kongshallen'.

Man kan også finde det 28. bind (blot som Poetiske Skrifter og med udgiveren Liebenberg som medforfatter) Her er overskrifterne i orden. Men hvor er alle de andre bind? Ingen kontrollerer åbenbart om man kun får et halvt leksikon eller en tyvendedel af de samle skrifter.

Søger man på: <Ludvig Holberg Peder Paars>, får man:

> **Peder Paars: et comisk heltedigt** Af Ludvig Holberg (baron), Adolph Engelbert Boye.

Boye er imidlertid udgiver og ordet 'Baron' er tilføjet i håndskrift på titelbladet. Der er kort sagt ingen kontrol af informationen i de enkelte poster.

Indtil videre ser det også ud til, at der er en stor skævvridning i forhold til engelsk-sprogede bøger. Desværre kan man ikke nogen steder se statistik over sprog. Men man kan 'snyde' ved at søge på et bestemt sprog og kun fuld tekst bøger ved kun at skrive et meget almindeligt forekommende ord på det pågældende sprog. En måling med mange svagheder, men den giver dog følgende vejledende billede:

1. Engelsk: 1.440.587
2. Fransk 246.600
3. Tysk 133.600
4. Dansk 14.958
5. Svensk 4.934
6. Russisk 3.052
7. Norsk 2.483

Det er som sagt kun de bøger, som er uden copyright, Google Books indeholder derfor rigtig mange bøger, de fleste, som stadig er omfattet af copyright og hvor fordelingen mellem de enkelte sprog sikkert er noget anderledes. Der er stadig rum for de europæiske sprog til at udfordre Google ved at gøre den trykte kulturarv digitalt tilgængelig.

For bibliotekerne er den digitale udfordring også kommet tæt på de sidste 15 år. Udlånet af fysiske bøger falder år for år, mens antallet af downloads af artikler stiger. Hvis også bøger, ligesom nu artikler fra tidsskrifter, bliver tilgængelige on-line i en helt anden størrelsesorden, vil det naturligvis ændre denne fordeling endnu mere. Hvordan dette vil ændre bibliotekernes udformning i fremtiden, er der i øjeblikket mange bud på. Men det vil næppe som forudsagt være bibliotekernes død, derimod en voldsom ændring i deres funktioner. Den elektroniske bogforsyning,

aktuelle sprog Alfabeter Anmeldelser arabisk Biblen bogstaver børn Danmark Dansk Dialekter engelsk esperanto Formidling fransk identitet konsonanter Medier modersmål Musik Navne norsk Ord ordbøger ordforråd oversættelse Plansprog religion romanske sprog russisk Sjov skriftsprog sprogdød Sproggeografi sprogkort Sprogpolitik sprogteknologi svensk truede sprog tv tyrkisk tysk Udtale Underholdning video vokaler

Arkiv

januar 2015
december 2014
november 2014
maj 2014
marts 2014
februar 2014
oktober 2013
august 2013
marts 2013
januar 2013
december 2012
november 2012
oktober 2012
september 2012
juli 2012
juni 2012
maj 2012
april 2012
marts 2012
februar 2012
januar 2012
december 2011
november 2011
oktober 2011
september 2011
august 2011
juli 2011
juni 2011
maj 2011
april 2011
marts 2011
februar 2011
januar 2011
december 2010
november 2010
oktober 2010
september 2010
juni 2010
maj 2010
april 2010
marts 2010
februar 2010
januar 2010
december 2009
november 2009
oktober 2009
september 2009
august 2009
juli 2009
juni 2009

Resources

Ethnologue: Languages of the World
Forvo – All the Words in the World. Pronounced.
LL-Map: Language and Location
Minority Rights Group
Omniglot. Writing Systems and Languages of the World
UNESCO Atlas of the World's Languages in Danger
World Atlas of Linguistic Structures (WALS)

Resurser

Bogstavlyd
Dansk sprognævn
Den danske ordbog
Dialekt.dk
dk.kultur.sprog
Korpus.dk
Nye ord i dansk på nettet (NOID)
Ordbog over det danske sprog
Ordnet. Dansk sprog i ordbøger og korpus
Sproget.dk
Svenska Akademien
ə (Schwa.dk)

som Google vil fremme eller monopolisere, alt efter hvordan man nu ser på det, er en god sikring af den bogmæssige kulturarv. Men det overflødiggør ikke biblioteker. Man skal stadig kunne komme tilbage til originalerne, både af videnskabelige grunde, men også fordi systemerne hele tiden skal kunne opfylde nye krav.

Man kan se en parallel i den revolution, som mikrofilmningen af aviserne medførte fra 1960erne og fremefter. I USA medførte det, at mange biblioteker smed deres indbundne årgange af aviserne ud. Det har ført til, at for flere lokalaviser eksisterer der ikke længere noget komplet originalt sæt. (7) Mikrofilmen er nu det nærmeste, vi kommer. Men det er en dårlig ide, fordi de første fotograferinger er meget dårligere end dem, som laves i dag, hvor man fx kan supplere med en OCR-scanning, så man kan lave elektronisk søgning i avisens tekstmasse.

Noget lignende ser vi i dag efter de mange on-line tidsskrifter er kommet: Mange biblioteker kasserer i dag deres indbundne årgange af tidsskrifter, som fylder på de dyre hylder. Men scanningen kan dels gøres bedre (fx er mange fysiske og medicinske artikler oprindeligt med farvebilag af grafer fx, men de er scannet i s/h) og hvordan sikrer man sig fx at der i Danmark findes ét sæt af alle de vigtigste tidsskrifter? Det er der ingen, der har overblik over. Så forskningsbibliotekerne vil i fremtidens samfund i en vis udstrækning få en vigtig funktion som steder for forskning og arkivering. De lokale biblioteker vil formentlig i højere grad blive steder, hvor man kan bruge medier, man ikke har adgang til via sin computer.

_____

[1] Robert Darnton: The Library in the New Age, *The New York Review of Books*, 12.6.2008.

[2] Fx kan man ikke trunkere i en Google-søgning og man kan kun lave en 'og'-søgning og ikke en 'eller'- og en 'ikke'-søgning, som er de muligheder den boolske logik giver i en normal databasesøgning.

[3] Gå ind fra google.com og vælge 'more'.

[4] For historien bag, se Robert J. Lackie: From *Google Print* to *Google Book Search*: The Controversial Initiative and Its Impact on Other Remarkable Digitization Projects, *The Reference Librarian* 2008/1, s. 35-53.

[5] Millie Jackson: Using Metadata to Discover the Buried Treasure in Google Book Search, *Journal of Library Administration* 2008/1/2, s. 165-73.

[6] *Information Today* oct. 2009, 7.

[7] Nicholson Baker: *Double fold. Libraries and the assault on paper*, New York 2001.

***Morten Thing**, dr. phil.*
Forskningsbibliotekar, Roskilde Universitetsbibliotek

maj 2009
april 2009
marts 2009

Læs også:

1. En e-bog er en ny slags bog! Apples chef, Steve Jobs, viser firmaets e-læser, iPad, frem. (Foto: Wikimedia) Bogen som medie går tilbage til skriftrullen og andre medier, som kan rumme en større mængde tekst. Det vi...
2. Hvad sker der med dansk skriftsprog på internettet? En stadig større del af de tekster som offentliggøres og læses på internettet, er skrevet af ikkeprofessionelle skribenter, og tekster af professionelle og ikkeprofessionelle optræder mellem hinanden. I forhold til...
3. En sprogskole i cyberspace Skolen i marts måned 2009 med landestedet og salgsautomaten til tilkøb af nye lektioner i forgrunden og parken i baggrunden Efter mange år som voksenunderviser på Studieskolen i København, begyndte...
4. Hjælpemidler i sprogundervisningen Elle mai s'arranger, quelques phrases sortent amende GoogleTranslate oversætter til fransk, men hvilket fransk? "Elle mai-être vrai, mais il peut aussi se tromper" er oversættelsen dansk-fransk i Google Translate af...

Tagget med: Amazon, Anmeldelser, Apple, aviser, bibliotek, bibliotek.dk, bog, boghistorie, bøger, citater, computer, copyright, digital, digitalisere, download, e-bøger, Ebrary, elektronisk, engelsk, Europa, forfatter, forlag, Frankfurt Bogmesse, Frankrig, fransk, Google, Google Books, Google Print, Google Scholar, Google World, Harvard, internet, Japan, jiddish, katalog, Kindle, kulturarv, kulturhistorie, Ludvig Holberg, Microsoft, mp3, nationalbibliotek, Norden, norsk, OCR, Oehlenschläger, opslagsværk, Oxford, Projekt Gutenberg, Projekt Runeberg, reklame, rettigheder, romanske sprog, russisk, sats, skrift, skriftkultur, sprogteknologi, Stanford, søgemaskiner, søgning, tegn, typografi, tysk, Tyskland, USA

1 kommentar

Morten Thing
20. november 2009 • 13:38

En ny aftale blev offentliggjort natten til lørdag den 14. november 2009. Den vigtigste ændring i forhold til den tidligere aftale er, at ikke-amerikanske bøger som udgangspunkt ikke længere er omfattet af forliget. Kun hvis en bog er registreret hos United States Copyright Office vil den være omfattet. Det drejer sig formentlig om meget få danske bøger. En af hovedindsigelserne mod den oprindelige forligstekst var netop, at den omfattede bøger udgivet af forlag og forfattere, som ikke havde haft nogen indflydelse på forliget. Den reviderede aftale omfatter således alene bøger registreret hos United States Copyright Office samt bøger udgivet før 5. januar 2009 i USA, Canada, Australien og England. Parterne har indgivet den reviderede forligsaftale for retten med en begæring om forhåndsgodkendelse af aftalen samt godkendelse af en supplerende meddelelse, der, hvis den godkendes, bliver sendt ud i starten af december 2009. Se hele den nye aftale samt yderligere information på: http://www.googlebooksettlement.com og http://www.authorsguild.org

Svar

Skriv en kommentar

Navn (kræves)

E-mail (kræves)

Hjemmeside

Send mig en e-mail når der kommer flere kommentarer.