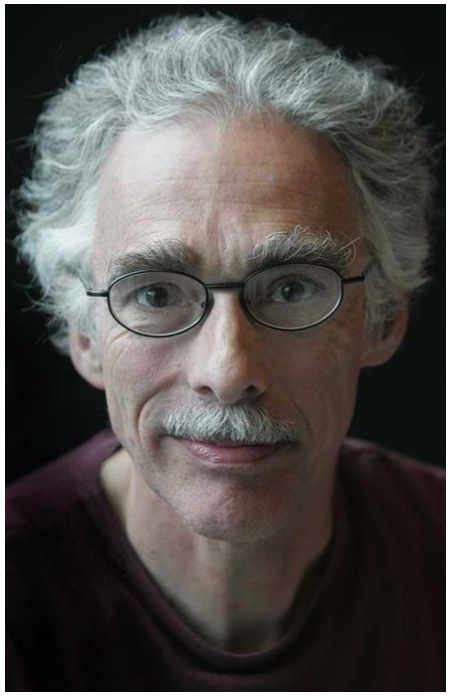


Eric Sieverts

De columns uit InformatieProfessional

1997-2015

Een onsje lever graag

"Weet jij het verschil tussen .....?"
Als een spreker zo begint, vertelt ons ingeslepen verwachtingspatroon dat vrijwel zeker een mop zal volgen. Toch was een serieuze vraag bedoeld. Andersom dan maar:
"Weet u de overeenkomsten tussen bedrukt papier en WWW?"

In onze maatschappij vervult bedrukt en beschreven papier primair de rol van informatiedrager. Dat is niets nieuws, daar bent u aan gewend. Ook aan het feit dat die informatie van alles zijn kan, een wetenschappelijk artikel, reclame van de slager op de hoek, een pornoblad, een dichtbundel, een liefdesbrief. Maar U realiseert zich wellicht ook dat voor het World Wide Web bijna exact hetzelfde geldt.

Toch zijn er ook verschillen. De aard van informatie op papier is meestal aan de vorm herkenbaar. Het flets gekleurde, rommelig opgemaakte A4tje van de slager ziet er anders uit dan een overdruk van een wetenschappelijk artikel. Een kloek deel van een encyclopedie valt van verre visueel te onderscheiden van het groezelige pornoblaadje. Op het web springen zulke verschillen veel minder in het oog. Elke pagina wordt in hetzelfde browser-jasje op het scherm geroepen. Pas bij zorgvuldig kijken zie je wat verschil.

Bezorging en distributie is ook anders. Een liefdesbrief zit in een (gekleurde en zacht geurende ?) geadresseerde envelop; de folder van de slager ligt samen met naar binnen gewaaide herfstbladeren op de deurmat; het sexblad moet opgehaald bij de sigarenwinkel of een ander detaillist van bedenkelijk allooi; de encyclopedie staat duidelijk herkenbaar in de boekenkast. Op het web komt alles door hetzelfde draadje onze PC binnen, wordt door dezelfde muisklik opgeroepen, met een gelijksoortig URL opgevraagd.

"Weet jij het verschil tussen .....?"
Als een spreker zo begint, vertelt ons ingeslepen verwachtingspatroon dat vrijwel zeker een mop zal volgen. Toch was een serieuze vraag bedoeld. Andersom dan maar:
"Weet u de overeenkomsten tussen bedrukt papier en WWW?"

In onze maatschappij vervult bedrukt en beschreven papier primair de rol van informatiedrager. Dat is niets nieuws, daar bent u aan gewend. Ook aan het feit dat die informatie van alles zijn kan, een wetenschappelijk artikel, reclame van de slager op de hoek, een pornoblad, een dichtbundel, een liefdesbrief. Maar U realiseert zich wellicht ook dat voor het World Wide Web bijna exact hetzelfde geldt.

Toch zijn er ook verschillen. De aard van informatie op papier is meestal aan de vorm herkenbaar. Het flets gekleurde, rommelig opgemaakte A4tje van de slager ziet er anders uit dan een overdruk van een wetenschappelijk artikel. Een kloek deel van een encyclopedie valt van verre visueel te onderscheiden van het groezelige pornoblaadje. Op het web springen zulke verschillen veel minder in het oog. Elke pagina wordt in hetzelfde browser-jasje op het scherm geroepen. Pas bij zorgvuldig kijken zie je wat verschil.

Bezorging en distributie is ook anders. Een liefdesbrief zit in een (gekleurde en zacht geurende ?) geadresseerde envelop; de folder van de slager ligt samen met naar binnen gewaaide herfstbladeren op de deurmat; het sexblad moet opgehaald bij de sigarenwinkel of een ander detaillist van bedenkelijk allooi; de encyclopedie staat duidelijk herkenbaar in de boekenkast. Op het web komt alles door hetzelfde draadje onze PC binnen, wordt door dezelfde muisklik opgeroepen, met een gelijksoortig URL opgevraagd.

Trekken en duwen

In Engels-sprekende landen placht ik altijd wat problemen te hebben met bordjes in warenhuizen en andere openbare gelegenheden, die aangeven in welke richting de deur scharniert. Bij de aansporing PUSH stond ik bijna onveranderlijk de deur naar me toe te rukken, terwijl het woordje PULL - veel gevaarlijker - me meestal niet op tijd liet afremmen. Die beide PU*-woorden lijken zoveel op elkaar, dat mijn brein ze niet tijdig uit elkaar kon houden.

Maar dat lijkt nu verleden tijd, want plotseling gonzen beide woorden dagelijks als modewoorden in ons vakjargon rond. Wie had twee jaar geleden al van PULL-media en PUSH-technologie gehoord? Nu wijdde zelfs NRC-Handelsblad er onlangs al een artikeltje aan. De termen suggereren nieuwe vergezichten en ongekende mogelijkheden. De teneur lijkt te zijn dat PULL-techniek eigenlijk ontzettend ouderwets is. Het is natuurlijk aardig dat je zelf het initiatief kunt nemen om te gaan zoeken, elke keer als je over een onderwerp iets nieuws wilt weten, maar eigenlijk is dat niet meer van deze tijd, met zijn robots en zijn "intelligent agents". Dus leve het PUSH-principe. Nieuwe informatie wordt vanzelf naar jou persoonlijk toegeduwd, of je het nu wilt of niet. Producten als PointCast en Marimba zorgen daar wel voor.

Het populaire PointCast stuurt via web-technologie automatisch nieuwe informatie naar je PC. Elke keer dat hij even staat niets te doen, vraagt die PC dat zelf op. Prachtig hoor. Ik kon uit wel 20 verschillende "kanalen" kiezen toen ik PointCast installeerde.

In Engels-sprekende landen placht ik altijd wat problemen te hebben met bordjes in warenhuizen en andere openbare gelegenheden, die aangeven in welke richting de deur scharniert. Bij de aansporing PUSH stond ik bijna onveranderlijk de deur naar me toe te rukken, terwijl het woordje PULL - veel gevaarlijker - me meestal niet op tijd liet afremmen. Die beide PU*-woorden lijken zoveel op elkaar, dat mijn brein ze niet tijdig uit elkaar kon houden.

Maar dat lijkt nu verleden tijd, want plotseling gonzen beide woorden dagelijks als modewoorden in ons vakjargon rond. Wie had twee jaar geleden al van PULL-media en PUSH-technologie gehoord? Nu wijdde zelfs NRC-Handelsblad er onlangs al een artikeltje aan. De termen suggereren nieuwe vergezichten en ongekende mogelijkheden. De teneur lijkt te zijn dat PULL-techniek eigenlijk ontzettend ouderwets is. Het is natuurlijk aardig dat je zelf het initiatief kunt nemen om te gaan zoeken, elke keer als je over een onderwerp iets nieuws wilt weten, maar eigenlijk is dat niet meer van deze tijd, met zijn robots en zijn "intelligent agents". Dus leve het PUSH-principe. Nieuwe informatie wordt vanzelf naar jou persoonlijk toegeduwd, of je het nu wilt of niet. Producten als PointCast en Marimba zorgen daar wel voor.

Het populaire PointCast stuurt via web-technologie automatisch nieuwe informatie naar je PC. Elke keer dat hij even staat niets te doen, vraagt die PC dat zelf op. Prachtig hoor. Ik kon uit wel 20 verschillende "kanalen" kiezen toen ik PointCast installeerde.

De wet van Wagenaar

De psycholoog Wagenaar, tegenwoordig rector-magnificus van de Leidse Universiteit, is vooral bekend als getuige-deskundige bij allerlei spraakmakende rechtszaken. Zijn specialisme is de kritische beschouwing van het selectieve en meestal zelfs ronduit onbetrouwbare geheugen van de mens, waar het in het verleden gedane waarnemingen betreft. Een ander, minstens zo interessant onderwerp waarmee Wagenaar zich bezig houdt zijn de psychologische effecten die ten grondslag liggen aan mede door menselijk handelen veroorzaakte rampen en ongelukken. De uit zijn waarnemingen afgeleide wetmatigheden met betrekking tot het risico-gedrag van mensen blijken in de praktijk op zo ruime schaal voor te komen, dat mijn vriendin het gemakshalve al over de Wet van Wagenaar heeft.

Die wet manifesteert zich inderdaad in allerlei vormen. Wanneer een automobilist veiligheidsriemen draagt, geeft hem dat een zoveel groter gevoel van veiligheid, dat hij geneigd is wat grotere risico's te nemen, zodat die veiligheidsriemen de veiligheid veel minder sterk verhogen dan men zou verwachten. Fietsers die 's avonds wél licht op hun fiets hebben, denken daardoor zoveel beter gezien te worden, dat ze minder defensief hoeven te rijden, met soortgelijk gevolg. Overigens moet dit niet als aansporing mijnerzijds worden opgevat om dan maar zonder licht of veiligheidsgordel te rijden.

Zonder dat het nu meteen om risico's gaat, zien we ook in ons dagelijks werk effecten die tot dezelfde klasse van wetmatigheden horen en die ik dus ook maar tot uitingen van de Wet van Wagenaar wil rekenen.

Ook U gebruikt natuurlijk E-mail. En ook u bent onmiddellijk bereid om iedereen die dat nog niet doet, de geweldige voordelen van dit communicatiemedium voor te schilderen. Zo veel sneller, zo veel makkelijker, kost zo veel minder tijd dan te moeten opbellen of brieven te moeten schrijven, in enveloppen te moeten stoppen en naar de brievenbus te moeten brengen.

Inderdaad, zo makkelijk dat ik dagelijks meer dan een uur bezig ben om al mijn E-mail af te handelen. Bijna meer tijd dan ik vroeger aan het schrijven en dichtplakken van brieven en aan het voeren van telefoontjes kwijt was. Veel van die mailtjes gaan namelijk over zaken waarvoor ik vroeger nooit moeite genomen zou hebben er een brief aan te wijden. Veel van die mailtjes komen van mensen die mij vroeger zeker nooit zouden hebben opgebeld. Al dat gemak en al die snelheid leveren uiteindelijk dus nauwelijks tijdwinst op. Maar, toch zou ik mijn E-mail voor geen geld meer willen missen.

En met de informatie op het web gaat het al niet anders. Je denkt nu zo veel makkelijker aan informatie te kunnen komen dan vroeger, toen je nog echt naar de bibliotheek moest om fysiek in tijdschriften, adresboeken of encyclopedieën te kunnen bladeren. Zelfs in het (onwaarschijnlijke) geval dat de verwachting over dat makkelijker vinden blijkt uit te komen, voorspel ik dat het je niet minder tijd zal kosten. Je zoekt veel langer door, omdat je ook steeds meer wilt hebben.

En of die verhoogde informatie-input dan tenminste de kwaliteit van het werk dat we afleveren ten goede komt, daarover doet de Wet van Wagenaar nu net geen uitspraak.

Openen en sluiten

De agenda van het december-nummer van ons blad vermeldde dat op 23 januari in Utrecht een symposium werd georganiseerd ter gelegenheid van de *opening* van de Elektronische Bibliotheek Utrecht. Bij het intikken van dit bericht was de typist van IP kennelijk nog met zijn gedachten bij een eerder artikel over de opening van zo'n bibliotheek in Groningen, want in werkelijkheid werd er in Utrecht iets gesloten!

Die opmerking vraagt natuurlijk om enige toelichting, want het sluiten van een elektronische bibliotheek is bijna net zo onmogelijk als het openen ervan. Dit zijn termen die aan fysieke bibliotheken zijn ontleend en niet meer van toepassing zijn op hun virtuele elektronische varianten. Ook wat Alex Klugkist vorig jaar in Groningen opende en waarvan hij in IP verslag deed, was eigenlijk geen elektronische bibliotheek. Het waren *alleen* een paar lokalen waarin wat oude en nieuwe PC's bij elkaar waren gezet. En dat maakt nog geen elektronische bibliotheek. Ja stil maar, ik weet natuurlijk best dat deze wat al te letterlijke opvatting mijn Groningse collega's vreselijk onrecht aandoet, want natuurlijk hebben ze daar ook wel degelijk een echte elektronische bibliotheek. Alleen kun je die niet echt zien en kon hij dus ook niet echt geopend worden.

Wat kan je dan wel met een elektronische bibliotheek, of het nu in Groningen of in Utrecht is? Die kan je vooral gebruiken. Want zo'n bibliotheek blijkt er op een gegeven moment gewoon te zijn. Niet in één klap, maar meestal heel geleidelijk aan, met steeds meer bronnen, steeds meer functies, steeds meer mogelijkheden en steeds vriendelijker interfaces. En dan niet alleen in een paar speciaal ingerichte lokalen. De elektronische bibliotheek is er pas echt als hij overal in de organisatie aanwezig is, liefst op ieders werkplek. En dat is natuurlijk wat er ook in Utrecht aan de hand was. Alleen komt zo'n bibliotheek er niet vanzelf.

Daarom werd op 23 januari in Utrecht *wel* degelijk iets afgesloten, namelijk een uitgebreid project onder de naam Elektronische Bibliotheek Utrecht.

Een project waaronder ongeveer 40 deelprojecten vielen (ik ben in de loop der tijd de precieze tel wat kwijt geraakt), die er op gericht waren ontwikkelingen op gang te zetten, kwesties uit te zoeken, prototypes te ontwikkelen, onderzoeken uit te voeren. En al die brokstukjes moesten bijdragen aan de verdere uitbouw van de elektronische bibliotheek zelf. Maar projecten zijn pas projecten als ze een eindige looptijd hebben (en dus worden afgesloten). Daarna zal nog moeten worden afgewacht welke daarvan voldoende levensvatbaar zijn om in de gewone routinematige praktijk te worden ingepast.

Gelukkig werden op 23 januari geen presentaties van alle 40 projecten gegeven. Integendeel. De langste bijdrage had zelfs niet rechtstreeks met het project te maken. Thijs Chanowski (achtereenvolgens bekend van Fabeltjeskrant - Freebase Text Retrieval - BSO - Origin Medialab - hoogleraarschap interface-kunde) mocht ruim een uur lang begeesterd uiteenzetten dat de huidige informatierevolutie heel andere denkwijzen van ons vraagt, ook al heeft op dit moment nog niemand een idee welke. Of het (mede door hem ontwikkelde) interface op het nieuwe bibliotheeksysteem van de Openbare Bibliotheek Eindhoven de richting wijst voor al onze informatie-zoekproblemen, vraag ik me nog wel een beetje af, maar dat het er heel schitterend en indrukwekkend uitziet zal niemand ontkennen.

Toch krijg ik op zwaarmoedige momenten wel eens het gevoel dat al die schitterende futuristische ideeën en systemen de aandacht wat afleiden van allerlei laag-bij-de-grondse praktische problemen. Is in uw bibliotheeksysteem het voor automatiseerders zo ontzettend saaie centennium-probleem bijvoorbeeld al opgelost? In het Utrechtse GEAC-systeem hadden we het 15 jaar geleden in elk geval al kunnen zien aankomen, toen een gebruiker met geboortejaar 1898 niet als lener bleek te kunnen worden ingeschreven. Als er iets snel moet worden afgesloten, is het dus wel het tijdperk van dergelijke oude bibliotheeksystemen.

[5] april 1998

Wie zijn nek uitsteekt, vangt veel wind

Ik ben in het bezit van een vouwfietsje - handig voor in de trein. Onlangs had ik er een nieuwe buitenband voor nodig. Vanwege de kennelijk ongebruikelijke maat moest de fietsenmaker die bestellen. De uiteindelijk afgeleverde band bleek bij monteren natuurlijk toch niet de goede maat te zijn, maar dat was de fietsenmaker niet aan zijn verstand te brengen. Toen hij bovendien weigerde de band terug te nemen en evenmin geld wilde teruggeven, ontspon zich een onverkwikkelijke discussie. Verklarend dat hij dan maar veel plezier aan die band moest beleven, heb ik toen met een snelle beweging de band om zijn nek gehangen. Dat laatste bleek niet zo verstandig, want deze fietsenmaker was niet alleen twintig jaar jonger, maar ook een kop groter met een tweemaal zo breed sportschoolpostuur. Enkele ogenblikken later stond ik dan ook blauw geschopt op straat, op de tast de resten van mijn bril bij elkaar zoekend.

Hiermee vergeleken gaan we in ons informatiewereldje een stuk vriendelijker met elkaar om. Daarbij doel ik nog niet eens op de ingezonden brief van Rob Klaverstijn en Jan van der Starre in het vorige nummer van *Informatie Professional*, waarin ze reageren op mijn kritische opinie over de Online Conferentie. Ook zij waren zo vriendelijk tegen me, dat ik onmiddellijk moest denken aan een bekende sketch met Wim Sonneveld als vader van een dochter in een moeilijke leeftijd. Als beste remedie tegen vader onwelgevallige vriendjes adviseerde hij zulke jongetjes hemelhoog het graf in te prijzen.

Nee, ik denk vooral aan collega's met wie je bij allerlei gelegenheden gesprekjes voert. Hester van der Kuur komt daarbij alleen collega's tegen die het helemaal eens zijn met haar visie op de IDM-opleiding, terwijl ze dat tegen mij nooit zeggen. De organisatoren van de Online Conferentie spreken louter tevreden bezoekers, terwijl ik in Rotterdam rondlopend voortdurend werd aangeklampt door mensen die verklaarden het helemaal met mijn kritiek op de conferentie eens te zijn. Het lijkt wel of men bang is klappen te krijgen als men een gespreksgenoot eens tegenspreekt.

Was ik toch in Rotterdam? Jazeker, zoals aangekondigd uitsluitend één dagje voor bezoek aan de tentoonstelling. Alleen vrees ik dat Van der Starre en Klaverstein mij (en ook een bus vol IDM-studenten en al die andere tentoonstellingsbezoekers) stiekem hebben meegeteld bij de tweeduizend congresgangers die zij noemden, want de uitgereikte deelnemerslijst bevatte er 'maar' ruim achthonderd.

Ik moet overigens bekennen blij te zijn dat de folder van de Dag van het Document nog niet was bezorgd, toen Jan en Rob hun reactie schreven. Die Dag is nog duurder dan de Online Conferentie en dan voor één dag. Dat ook mijn naam bij het rijtje bobo's staat dat voor het programma verantwoordelijk is, was natuurlijk een kans voor open doel geweest. Maar wees gerust, ook die hoge prijs zit me dwars. Een aantal Dagen van het Document geleden mocht je voor een aanzienlijk lager bedrag nog met zijn tweeën komen. (Al weet ik ook dat sommige commerciële organisaties wel duizend gulden per dag vragen). Voor VOGIN en NVB heb ik zelf themadagen of -middagen mee georganiseerd, waarvan de programma's - wellicht onbescheiden opgemerkt - niet voor deze congressen onderdeden. Maar de discrepantie tussen toen gemaakte onkosten en wat tegenwoordig 'echte' conferenties kosten, blijft bevreemden.

Collega-redacteur Paul Nieuwenhuysen van *Informatie Professional* herinnerde er terecht aan dat onze Online Conferentie ooit was opgezet als low-budget alternatief voor wie niet naar de dure conferentie in Londen kon of mocht. Waar gaat al dat geld nu dan heen? Nauwelijks naar sprekers, niet naar organiserende non-profit organisaties, niet naar programmacommissies. Kennelijk weten verhuurders van accommodaties (zelfs als ze een hightech congres niet meer dan modemlijntjes kunnen bieden) munt te slaan uit de overspannen congres- en symposiummarkt. Daar mogen we best eens kritisch over zijn.

Intussen heb ik mijn eerste sportschool-training afgesproken. Fysieke weerbaarheid blijkt nodig. De volgende misstand dient zich al weer aan: mijn laatste IBL-aanvraag. Waarom kreeg ik die van de week terug met de mededeling dat het betreffende tijdschrift bij de eerste bibliotheek niet aanwezig was, bij de tweede naar de binder en dat het DUS niet leverbaar was, terwijl dat tijdschrift volgens de NCC bij nog minstens vijf andere bibliotheken aanwezig moet zijn? Denken ze soms dat ik die aanvraag voor de lol heb ingestuurd, om te pesten, dat ik dat artikel eigenlijk helemaal niet nodig heb?

Pas maar op, ik ben er klaar voor. Dat fietsbandje bleek ik ineens weer, samen met mijn brillenglazen in mijn hand te hebben. Wie is de volgende die hem om zijn nek krijgt?

[6] mei 1998

Yahoo! won

Yahoo won! Met die uitroep begon één van de lezingen (van Deborah Lynn Wiley) op een congres over retrieval software enkele weken geleden in Boston. Alsof het ging om een wedstrijd tussen de Boston Red Socks en de Chicago Bulls. De spreker bedoelde daarmee echter dat YAHOO! volgens sommige onderzoeken onder Internet-gebruikers een populairder zoekhulpmiddel is dan "echte" zoekmachines als AltaVista of HotBot. Hoewel op die ongenuanceerde uitspraak wel wat valt af te dingen, viel een soortgelijke teneur ook in andere lezingen te beluisteren. Dat had dan niet specifiek betrekking op YAHOO en zelfs niet op de ontsluiting van het Internet, maar op de toepassing van retrieval-technieken in het algemeen.

De euforie van de mooie full-text zoekmogelijkheden, zelfs van de huidige niet-Booleaans methoden met vectormodellen, fuzzy-searching, probabilistische methoden, relevance ranking, natuurlijke taal-technieken en wat dies meer zij, begint wat weg te ebben. Ondanks de geleidelijk betere prestaties van retrieval software bij de jaarlijkse TREC-competitie. Bij deze wedstrijd tussen full-text retrieval-programma's worden centraal opgegeven zoekvragen op een zelfde corpus van vele gigabytes aan veelsoortige full-text data losgelaten. De laatste zes jaar is de behaalde precisie (bij vaste recall) gemiddeld bijna twee keer zo goed geworden (al is die nog altijd niet echt geweldig). Maar het feit dat de zoekacties door de makers van de programma's worden uitgevoerd en niet door toevallige gebruikers, maakt dat die resultaten eigenlijk weinig relatie met de dagelijkse zoekpraktijk hebben.

Mede daardoor lijkt men weer meer oog te krijgen voor het toepassen van klassieke bibliotheek-achtige technieken. En daarvan is YAHOO natuurlijk een aardig voorbeeld. Daar is een systematische indeling opgezet (wat wij als informatiespecialisten ook van de kwaliteit van die indeling mogen vinden). Daar worden door YAHOO's bibliothecarissen Internet-bronnen gecollectioneerd (wat wij ook van het collectiebeleid mogen vinden). En dat ook ongevraagd opgestuurd materiaal wordt opgenomen (daar: door anderen aangemelde URL's), is in de bibliotheekwereld evenmin ongewoon.

Yahoo won! Met die uitroep begon één van de lezingen op een congres over retrieval software enkele weken geleden in Boston. Alsof het ging om een wedstrijd tussen de Boston Red Socks en de Chicago Bulls. De spreker bedoelde daarmee echter dat YAHOO! volgens sommige onderzoeken onder Internet-gebruikers een populairder zoekhulpmiddel is dan "echte" zoekmachines als AltaVista of HotBot. Hoewel op die ongenuanceerde uitspraak wel wat valt af te dingen, viel een soortgelijke teneur ook in andere lezingen te beluisteren. Dat had dan niet specifiek betrekking op YAHOO en zelfs niet op de ontsluiting van het Internet, maar op de toepassing van retrieval-technieken in het algemeen.

De euforie van de mooie full-text zoekmogelijkheden, zelfs van de huidige niet-Booleaans methoden met vectormodellen, fuzzy-searching, probabilistische methoden, relevance ranking, natuurlijke taal-technieken en wat dies meer zij, begint wat weg te ebben. Ondanks de geleidelijk betere prestaties van retrieval software bij de jaarlijkse TREC-competitie. Bij deze wedstrijd tussen full-text retrieval-programma's worden centraal opgegeven zoekvragen op een zelfde corpus van vele gigabytes aan veelsoortige full-text data losgelaten. De laatste zes jaar is de behaalde precisie (bij vaste recall) gemiddeld bijna twee keer zo goed geworden (al is die nog altijd niet echt geweldig). Maar het feit dat de zoekacties door de makers van de programma's worden uitgevoerd en niet door toevallige gebruikers, maakt dat die resultaten eigenlijk weinig relatie met de dagelijkse zoekpraktijk hebben.

Mede daardoor lijkt men weer meer oog te krijgen voor het toepassen van klassieke bibliotheek-achtige technieken. En daarvan is YAHOO natuurlijk een aardig voorbeeld. Daar is een systematische indeling opgezet (wat wij als informatiespecialisten ook van de kwaliteit van die indeling mogen vinden). Daar worden door YAHOO's bibliothecarissen Internet-bronnen gecollectioneerd (wat wij ook van het collectiebeleid mogen vinden). En dat ook ongevraagd opgestuurd materiaal wordt opgenomen (daar: door anderen aangemelde URL's), is in de bibliotheekwereld evenmin ongewoon.

* Deborah Lynn Wiley / Beyond Information Retrieval: Ways to Provide Content in Context (1998) - http://www.cs.uml.edu/~haim/teaching/iws/tirsaa/sources/Beyond_IR_Content_in_Context.html

De wet van de remmende voorsprong

Omdat we Frankrijk af hadden, zijn we de afgelopen zomervakantie maar eens aan Spanje begonnen. Netjes systematisch van bovenaf. Alle vooroordelen die nog over Spanje mochten leven, werden meteen gelogenstraft. WC's op de campings waren schitterend en bovendien onveranderlijk voorzien van rollen toiletpapier, zodat je niet zoals in Frankrijk de eigen rol onder de kin hoefde klemmen. Bovendien bleek het eerste de beste dal waar we terecht kwamen, hoog in de Pyreneeën, al zijn eigen URL te hebben. Wie de mailtjes naar het bijbehorende mail-adres zou gaan beantwoorden, werd niet helemaal duidelijk. Niet dat dat erg was, want in mijn tent heb ik toch geen communicatievoorzieningen. Geen notebook, zoals in de grote caravan verderop, geen GSM, geen modem. Ik ga al bijna 20 jaar online, maar nog niet tijdens mijn vakantie. Zou dat de wet van de remmende voorsprong zijn?

Foldertjes van mijn vakantiebestemmingen in Frankrijk hadden mij trouwens nog nooit op bijbehorende web-pagina's gewezen. En op de eigen web-pagina van een ander Pyreneeëndal, de Cerdagne, met zowel een Frans als een Spaans gedeelte, blijken alleen de Spaanse plaatsen doorklikmogelijkheid te bieden. Toch is Frankrijk jarenlang het voorbeeld en de voorloper geweest voor grootschalige elektronische informatievoorziening via hun Minitel-systeem. Op basis van de hier in Nederland (volgens mij trouwens terecht) altijd zo versmade Viditel-techniek. Zou Frankrijk op dit terrein ook last hebben van remmende voorsprong?

Nog meer voorbeelden?

Anderhalf jaar geleden schreef ik in deze rubriek iets over push- en pull-techniek. Toen was PointCast HET grote voorbeeld van de nieuwe push-techniek. Maar als ik recente berichten in de kranten mag geloven, is PointCast nu al op een haartje na failliet. En America Online. Dat was tien jaar geleden de eerste op het algemene publiek gerichte grote online host. En nu?

..... Pardon, verkeerd voorbeeld. Die wet van de remmende voorsprong lijkt dus toch niet altijd op te gaan. Is het dan nog wel een wet? In de wiskunde en de natuurkunde is één tegenvoorbeeld al genoeg om een wet te ontkrachten. Dus misschien geen wet, maar wel iets dat we elkaar voortdurend lijken aan te praten. Dan wordt het vanzelf wel een voorspelling die zichzelf vervult. (En is dat niet weer een andere wetmatigheid?)

Toch weet ik een uitstekend voorbeeld waarop die zogenaamde wet van de remmende voorsprong wel van toepassing blijft: op mijn eigen notebook. Ik voel me net de man in de radioreclame, die stinkjaloers is op het nieuwe blitse A5 notebookje van zijn baas. Maar voor hij er ook zo een krijgt, moet hij eerst zijn oude opmaken. En, van hetzelfde merk, wil die maar niet stuk. Bij mij ook niet (ook al is die van een ander merk). Dus heb ik nog steeds zwart-wit, Windows-3.11, 50 MHz en "maar" 250 MB harde schijf. Er is gelukkig wel een geruststelling: als ik nu zo'n mooi nieuw hebbedingetje (met Intel Pentium-II inside) koop, ben ik over 3 jaar toch weer opnieuw tot achterstand afgeremd.

[8] april 1999

John's videoclip-syndroom

Ik moet eindelijk maar eens bekennen dat ik eigenlijk heel ouderwets ben. Al heb ik sinds twee jaar televisie, kijken doe ik nog altijd niet veel. Geen GTST, weinig gezap en al helemaal geen videoclips. Daar moest ik aan denken toen ik drie maanden geleden uitspraken van Hoos Blotkamp, directeur van het filmmuseum, in ons blad tegenkwam. Zij maakt zich zorgen over de oppervlakkigheid van het huidige informatietijdperk. "Al zappend kunnen [scholieren en studenten] zich [..] geestelijk voeden met flarden tekst en beeld, totdat hun hoofden zo leeg zijn als een volle prullenbak". Krek zo is het maar net, dacht de ouwe sok.

Een maand later moest ik daar nog veel sterker aan denken. Toen werd ons door John Mackenzie Owen het einde van het document aangezegd. Als die aanzegging zich tot het gedrukte document beperkt zou hebben, had ik instemmend geknikt. Mackenzie Owen's bewering ging echter veel verder. Hij signaleert een fragmentarisering van informatie en meent dat er "uiteindelijk slechts bouwstenen bestaan waarmee iedere lezer een eigen inhoud [..] opbouwt". Terwijl iedereen de mond vol heeft van kennis-management, van veredeling van informatie tot kennis, dreigt die kennis zo weer te worden gereduceerd tot losse stukjes informatie zonder samenhang en die informatie tot data en losse bytes.

Eerder benadrukte Blotkamp het belang van reflectie. Dat bij de doorgifte van denkbeelden volledigheid, helderheid en eenduidig taalgebruik van belang zijn. Dat is nu juist wat gaat ontbreken, als ieder zijn eigen informatiemandje zelf uit losse fragmenten moet gaan vullen. Geen zorgvuldig opgebouwde betogen meer, geen logische redeneringen die tot onontkoombare conclusies leiden, kortom het einde van wetenschap en cultuur.

Het klinkt misschien vreemd deze opvatting te horen van een docent die zijn studenten al vele jaren enthousiast de verworvenheden van het hypertext-concept uitlegt. De voordelen losse stukken informatie via hyperlinks te kunnen koppelen. De vrijheid om naar eigen goeddunken zelf de gewenste weg door een hyperdocument te kiezen. Hyper*document*?

Inderdaad document! Goede hypertext- of hypermedia-systemen moet je wel degelijk documenten noemen. Zulke systemen zijn namelijk pas goed, als ze zorgvuldig zijn gecomponeerd. Alleen als al die hyperlinks goed overdacht zijn aangebracht. Alleen als de auteur, de samensteller, de structureerder, en daarmee de ontsluiter van zo aangeboden informatie, daarin structuur, logica en verband heeft aangebracht. Een taak, overigens, met een sterk informatie-specialistische dimensie. Een willekeurig ratjetoe van losse fragmenten is heel wat anders. Dat is waar Hoos Blotkamp bang voor is. Dat lijkt de toekomst die John Mackenzie Owen voor ons in het verschiet ziet.

Natuurlijk koester ik niet de illusie dat mensheid, maatschappij en media zich iets van de mening van Eric Sieverts zullen aantrekken. Maar ik heb wel hoop dat de videoclip-benadering, de atomisering van informatie, kennis, cultuur en wetenschap een modeverschijnsel is. Zoals de management-modes die elkaar in een bonte cyclische stoet lijken af te wisselen. Zoals de damesmode, waar de hoogte van de rokrand een fraai golvende sinus-functie tegen de tijd beschrijft. Maar eigenlijk geloof ik ook helemaal niet dat John gelijk heeft.

Als hij schrijft dat "het gebied van de gepubliceerde documenten [..] steeds meer een marginaal verschijnsel [wordt], het aandeel hiervan in het geheel van de digitale informatie al bijna verwaarloosbaar klein [is]", dan is dat gewoon niet waar. Er schijnen ruim 500 miljoen web-pagina's te zijn. Dat is nog altijd aanzienlijk minder dan grote hosts aan gepubliceerde documenten in hun - inderdaad wat minder frequent geraadpleegde - databases aanbieden. En ook onder die 500 miljoen webpagina's zijn nog een heleboel "documenten".

Misschien moeten we van de dames- (en heren-)mode leren dat alle stijlen best naast elkaar kunnen voorkomen. Punk naast mantelpak. Gestructureerde, uit zorgvuldige redeneringen opgebouwde documenten, naast door elk individu naar eigen goeddunken uit toevallige losse data-fragmenten samengestelde hyper-info-scripten.

De wetten van Garfield

Even een testje: "wat is je eerste associatie bij het horen van de naam Garfield?
Kat, strip, comic, ... of *ISI, citatieindex, Eugene, Journal Citation Reports, ...*
Ik voorzie dat de antwoorden in twee groepen van dit soort uiteen zullen vallen. Daaruit kun je meteen afleiden wie jeugdige informatieprofessionals zijn en wie al in het vak zijn vergrijsd. Die laatsten plachten te smullen van de stukjes van Eugene Garfield, oprichter van ISI, voorin de (nog papieren) Current Contents, en hebben hem misschien nog wel zelf met zijn Einstein-haardos en zijn sjofele aktetasje op de eerste On Line congressen zien rondlopen.

Gebeurt er nog wel iets met zijn oorspronkelijke ideeën over citatiegedrag, anders dan misbruik door boekhouders voor verdeling van onderzoeksgeld? In wetenschappelijke artikelen opgenomen literatuurreferenties laten dienen als gratis extra inhoudelijke ontsluiting voor retrieval van echt relevante publicaties. Meten wat toonaangevende tijdschriften zijn door middel van datzelfde citatiegedrag. Het inzicht dat wetenschappelijk publiceren dankzij opgenomen literatuurreferenties een zelf-organiserend systeem is. Statistische analyse door de computer wie wie citeert en welke artikelen vaak samen in literatuurlijstjes voorkomen, levert een dynamisch beeld hoe de wetenschap in disciplines en deelgebiedjes is opgedeeld, hoe die in de tijd evolueren en welke dwarsverbanden er bestaan.

Die "wetten" van Eugene Garfield lijken herontdekt te worden op het web. Dat je hyperlinks als een variant op literatuurreferenties kunt beschouwen en je met bepaalde zoekmachines dus ook kunt citatiezoeken op het web, kwam in onze WWW-rubriek al eens aan de orde. Maar ook van collectiever citeergedrag wordt gebruik gemaakt. Zo beschouwt Google het aantal keren dat via hyperlinks naar een web-pagina wordt verwezen, als maat voor kwaliteit en daarmee voor relevantie.

In het juni-nummer van Scientific American was een heel artikel gewijd aan het gebruik van dergelijke citatierelaties als mogelijke oplossing voor retrieval-problemen op het web. Het web groeit zo snel dat zoekmachines die groei niet meer kunnen bijhouden. Uit een jaarlijks onderzoek van de NEC-corporation kwam als schatting dat afgelopen voorjaar het aantal web-pagina's tot ruim 800 miljoen was gegroeid. Zelfs de volledigste zoekmachine - volgens datzelfde onderzoek nu Northern Light - had daarvan maar 16% geïndexeerd (Nature, 8 juli 1999, p.107)! De beste van vorig jaar, HotBot, had toen nog een dekking van 34% van de berekende 320 miljoen pagina's. Als we anderzijds bedenken hoeveel wel geïndexeerde pagina's trivialiteiten, geleuter, leugens en verdubbelingen bevatten, wordt duidelijk dat er behoefte is aan andere retrieval-methoden die meer op kwaliteit dan op kwantiteit zijn gericht.

De auteurs van het stuk in Scientific American, onderzoekers bij IBM, doen dat door twee soorten belangrijke web-pagina's te onderscheiden: *authorities* en *hubs*. *Authorities* zijn pagina's met belangrijke primaire informatie. Dat ze belangrijk en primair zijn, wordt afgeleid uit het feit dat er vanuit veel andere pagina's naar wordt verwezen. *Hubs* zijn belangrijke verwijspagina's. Dat ze belangrijk zijn en vooral een verwijsfunctie hebben, wordt afgeleid uit het feit dat ze naar veel *authorities* verwijzen. Een *authority* kun je dus ook definiëren als een pagina waar veel *hubs* naar verwijzen. Hoewel dat een cirkelredenering lijkt, blijk je door iteratieve rekenprocessen toch tot kwaliteitsoordelen over zowel *authorities* als *hubs* te kunnen komen. Bij elke zoekactie kan de computer uitrekenen wat de beste *hubs* en *authorities* zijn die aan je vraag voldoen, zodat die als relevantste het eerst getoond kunnen worden. Dat dat nu nog wat veel rekentijd vergt, is ongetwijfeld een voorbijgaand probleem.

Dus opnieuw betere retrieval-resultaten op basis van citatiegedrag. Lang leve Garfield - en dan bedoel ik dus Eugene!

Zoekers rommelen maar wat aan

"*Zoekmachines op Internet rommelen maar wat aan*". Onder die titel bracht de Automatisering Gids van 7 april een journalistieke weergave van het ook in IP gerapporteerde onderzoek van Wouter Mettrop en de zijnen naar het onbetrouwbare zoekgedrag van zoekmachines op het web. Enkele beheerders van zoekmachines - Ilse, AltaVista - was om commentaar gevraagd. Zij gaven toe dat zoekacties inderdaad niet altijd helemaal worden afgemaakt of niet op de hele - gedistribueerd opgeslagen - index worden losgelaten, om de responstijd voor alle duizenden gelijktijdige zoekers acceptabel te houden. Dat wordt belangrijker geacht dan volledigheid, betrouwbaarheid en reproducibiliteit. Je vindt immers altijd nog wel iets!

Voor een grote meerderheid van Internetgebruikers kan dat inderdaad gelden. Maar voor professionele zoekers is dat nauwelijks acceptabel. Die weten hoe ze zoeken moeten. Toch?

Enige tijd geleden had ik een week proeftoegang tot de Nederlandse Persdatabank, via het web. Een aardig product. En een goede gelegenheid om dat ene artikel op te zoeken, dat ik destijds zo keurig uit de NRC geknipt had. Een artikel - toepasselijk - van Karel Knip, dat - even toepasselijk - gewijd was aan het vorig jaar in Nature gepubliceerde onderzoek naar de grootte van zoekmachines op het web. Het onderzoek dat opleverde dat zelfs de grootste zoekmachines maar een klein gedeelte (16%) van het geschatte totaal van 800 miljoen webpagina's doorzochten. Helaas was Karel's knipsel al weken zoek. Ergens in één van de random-access stapeltjes op mijn bureau lag het, maar in welk?

Dat kan de computer beter:
"**www AND zoekmachines AND grootte**"

Hoe kan dat? Het zit er niet bij. En waarom wel een artikel over het broeikaseffect, toevallig ook van KK? Even de volledige tekst bekijken. Ach ja, KK is zelf een verwoed web-zoeker en rapporteert vaak hoe hij iets over zijn onderwerp op Internet heeft kunnen vinden.

Zoekvraag dus niet goed? Misschien niet GROOTTE van zoekmachines, maar GEDEELTE van het web dat ze doorzoeken:
"**www AND zoekmachines AND gedeelte**"

Eén artikel, interessant, maar niet wat ik zoek.
Nieuwe poging:
"**www AND zoekmachines AND fractie**"

Helemaal niets.
"**www AND zoekmachines AND aantal**"

Zeven treffers, maar niet de goede.
Zelfde combinaties met "**internet**" in plaats van "**www**".
Allerlei resultaten, soms weer Karel Knip, maar niet die ene.

Ja, stil maar. Ik weet dat ik na de eerste misser meteen op auteursnaam had kunnen zoeken. Maar het was mijn eer te na die makkelijke weg te bewandelen. Op woorden uit de inhoud moet je het toch ook kunnen vinden. En inderdaad: het artikel zat er in. En al die gebruikte zoektermen die zo ontzettend vanzelfsprekend waren, waarvan je je absoluut niet kon voorstellen dat ze NIET in het artikel zouden voorkomen, die bleken er .... inderdaad NIET in voor te komen.
Met "**www AND zoekmachines AND deel**" was het wel te vinden.

Die ervaring maakte mij heel bescheiden. Ik kon me ineens heel goed verplaatsen in de gevoelens van al die deelnemers aan de VOGIN-cursus, die een zoekopdracht - met wedstrijdelement - in een bibliografische database krijgen. Die elke cursus opnieuw niet meer dan 2 tot 20 treffers krijgen, waarna de docenten tonen hoe ze door slim gebruik van gecontroleerde ontsluiting minstens 200 relevante artikelen hadden kunnen vinden.

Inderdaad, al die zoekers - ik niet uitgezonderd - rommelen maar wat aan.

Informatie overload

Novecento is een beroemde film van Bertolucci over de nu afgelopen eeuw. Ook al is hij al in 1976 gemaakt en al gaat hij eigenlijk alleen over de eerste helft van de eeuw. Die Italiaanse eeuw, de jaren negenhonderd - de duizend wordt voor het gemak weggelaten, zoals wij over de "sixties" spreken - liep dus in elk geval een jaar geleden al af. Voor scherpslijpers zoals ik, is dat met onze twintigste eeuw pas nu, met het afsluiten van het jaar 2000, het geval. Italiaans of Nederlands, de nieuwe eeuw en het nieuw millennium zijn nu voor iedereen begonnen. Eindelijk mag ik vooruit kijken, maar zeker geen millennium en eigenlijk nauwelijks een hele eeuw. Er is immers weinig kans dat iemand van de lezers van vandaag mijn uitspraken aan het eind van deze eeuw kan controleren.

Tot dat vooruitkijken werd ik gebracht door een recent onderzoek van twee Amerikanen, Peter Lyman en Hal Varian. Zij hebben uitgerekend hoeveel informatie we met z'n allen per jaar produceren, in enigerlei vorm, dus ook foto's, video en papier. In hun speciale extreem korte samenvatting, bedoeld voor lijders aan "heavy information overload", melden ze dat dat 1,5 exabyte per jaar is, als je het allemaal in de computer zou opslaan. Omdat ik nog nooit van exabytes gehoord had, was ik blij dat ze er zelf bij zeiden dat dat 1,5 miljard gigabyte is. In hun iets uitgebreider rapportage voor lezers met "moderate information overload" becijferen ze dat ruim 90 procent van die informatie al metterdaad in digitale vorm geproduceerd wordt.

Het onderzoeksteam van Lyman en Varian heeft ook berekend dat we onze productie elk jaar verdubbelen. Als dat altijd al zo geweest was, zou dat betekenen dat we het afgelopen jaar net zo veel informatie geproduceerd hebben als de totale mensheid in alle afgelopen eeuwen en millennia bij elkaar. Dat het altijd zo geweest is, geloof ik niet, want een verdubbeling per jaar is nog heel andere koek dan de onderzoeksresultaten van Derek de Solla Price uit de jaren zestig - dat decennium dat in 1969 eindigde. Hij rekende uit dat de productie aan officieel gepubliceerde informatie sinds de zeventiende eeuw - ja inderdaad, de "seicento" - al bijna drie eeuwen lang elke 15 jaar verdubbelt. Dat werd beschouwd als een ongekend lange ononderbroken exponentiële groei, zoals van geen enkel ander natuurverschijnsel bekend was. Als het nu ieder jaar verdubbelt, kan dat niet anders dan een tijdelijke versnelling zijn.

Denk maar aan het sprookje van de uitvinder van het schaakspel. Als beloning vroeg hij aan de koning één graankorrel op het eerste veld van het schaakbord, en op elk volgend veld het dubbele. De met weinig rekenkundig inzicht begiftigde koning doorzag niet dat daarmee een hoeveelheid graan gemoeid was, die de volledige jaarproductie van de hele aarde ver te boven ging. Heeft het schaakbord maar 64 velden, een eeuw heeft 100 jaren. Een verdubbeling elk jaar zou maken dat onze productie toeneemt met een factor 10 tot de 30ste macht (een 10 met 30 nullen, nog net geen googol, de 10 met 100 nullen waar de zoekmachine Google naar vernoemd is). Hoe lang de wet van Moore over de verdubbeling van de capaciteit van onze computers en onze computergeheugens ook al geldig is, met minder dan één atoom zullen we de komende eeuw nog niet toekunnen voor het opslaan van één bit informatie. Om onze 10 tot de 49ste bits in 2099 te kunnen opslaan, zou ten minste 10 procent nodig zijn van alle atomen waaruit de hele aardbol bestaat.

Mijn voorspelling dat die jaarlijkse verdubbeling ruim voor het eind van de net begonnen eeuw zal zijn afgevlakt, is dus nauwelijks riskant te noemen, hoe goede compressietechnieken er nog ontwikkeld zullen worden. En dan heb ik nog niet eens voorgerekend dat anders uw achterkleinkinderen en de hele verdere wereldbevolking meer dan een miljard maal een miljard exabyte per persoon per jaar zouden moeten produceren. Nee mijn voorspelling voor 2099 zit wel goed: dit zal in de eenentwintigste eeuw niet gebeuren.

Zouden de Italianen trouwens zelf al weten wat op novecento gaat volgen, hoe zij hun in 2099 aflopende variant van de eenentwintigste eeuw gaan noemen, hoe over 75 jaar de film van de nieuwe Bertolucci moet gaan heten? Ik zou in elk geval niet weten hoe ik er in een zoekmachine naar zou moeten zoeken. Hoe zoekmachines over 100 jaar tegen al die exabytes opgewassen zullen zijn, is natuurlijk nog weer een heel andere vraag.

* Lyman & Varian / How Much Information? 2000 - http://groups.ischool.berkeley.edu/archive/how-much-info/

Geen overload maar inflatie

Begin dit jaar slaakte ik verzuchtingen over informatie overload, daartoe aangezet door resultaten van Amerikaans onderzoek (zie vorige column). De voor een column uitgetrokken kilobytes lieten toen geen ruimte voor enige kritische reflectie. Toch was daar wel reden toe. Is het alleen maar een uiting van informatie overload, die 1,5 exabyte - anderhalf miljard gigabyte - die we volgens onderzoekers Lyman en Varian in 1999 hebben geproduceerd? En de 3 exabyte die het volgens hun extrapolatie in 2000 moeten zijn geweest? Nee natuurlijk niet. Bits en bytes zijn een heel verkeerde maat om informatie mee te meten.

Twintig jaar geleden, mijn eerste ervaringen met online zoeken nog heel vers, haalde ik met mijn 300 baud modem informatie over wetenschappelijke artikelen op. Gemiddeld 2 kB per artikel. In een minuut waren die gegevens binnen. Inderdaad alleen maar titels, abstracts en trefwoorden, maar toch al de essentiële informatie om te weten waarover die artikelen gingen. Tien jaar geleden waren heel wat artikelen full-text beschikbaar. Met 2400 baud kwam die 30 kB platte tekst, kale ASCII voor acht A4-tjes wetenschap, in net acceptabele tijd - twee minuten - binnen. Tabellen zagen er nog niet mooi uit en grafische illustratie van het geschrevene ontbrak. Word-bestanden of PDF-files maken dat heel wat mooier. Maar die acht pagina's worden dan al gauw 100 tot 400 kB. En als een scanner is gebruikt om papieren artikelen te digitaliseren, kom je voor datzelfde artikel in compressed TIFF, met voldoende resolutie om ook een grafiek te kunnen aflezen, al snel tot 1 MB.

Is dat het einde? Voor wetenschappelijke informatie op dit moment misschien nog wel. Maar met de komst van multimedia, kunnen we tekst ook laten voorlezen, zoals bij het radiojournaal via internet. Over diezelfde acht A4-tjes doet een spreker een minuut of twintig, ofwel 3 MB streaming audio. Echt interessant wordt het natuurlijk pas als je ook kunt zien hoe een wetenschapper - hakkelend en wel - zijn tekst voorleest, overheadprojector, imposante boekenkast of borrelende reageerbuis op de achtergrond. Met streaming video via internet komt die twintig minuten bij voldoende bandbreedte in twintig minuten binnen.

Nog lang niet schermvullend, maar toch al 200 MB - voor datzelfde artikel, voor diezelfde informatie. Dat is nog eens inflatie. In twintig jaar van 2 kB naar 200 MB. Zo zou een broodje van fl. 2,- toen, nu 2 ton moeten kosten.

Toch vormen plaatjes, geluid en video veel van de "informatie" waarmee Lyman en Varian tot hun 1,5 exabyte kwamen. Geen informatie dus, maar bits en bytes waarin de echte informatie in steeds verdunder vorm verstopt zit. Dus eigenlijk vooral begripsinflatie. Dezelfde soort begripsinflatie waarmee sommigen informatie voor kennis aanzien. Maar dat is weer een ander onderwerp.

Ik zie wel een andere parallel. Met veel fanfare is onlangs aangekondigd dat het menselijk genoom bijna geheel ontrafeld is. Ons DNA blijkt 3 miljard base-paren te bevatten - ja van de zuren en de basen uit de scheikunde-les. Voor elke base biedt de natuur keuze uit maar vier soorten en elke soort vormt een vast paar met maar één andere. Met 3 GB is de hele mens dus vastgelegd. Past makkelijk op de harde schijf van een enkele PC. Is dát dan wel informatie? Nee, ook alleen nog maar gegevens, want intussen heeft men ook geschat dat in die 3 miljard base-paren maar 30.000 genen verborgen liggen. Nauwelijks meer dan van een rondworm of de zandraket. Toch zijn die genen de echte dragers van onze genetische informatie.

Als met die paar gigabyte en dat kleine beetje informatie over onze genen al een heel mens gecodeerd kan worden, wat voor interessants zouden we met 1,5 exabyte dan wel niet kunnen doen. Daar zien we onze informatie-inflatie pas. Die exabytes worden nu gebruikt voor vakantiekiekjes, voor al die informatie die vroeger nooit verder kwam dan de verwaaiende kreten van een verjaarspartijtje, dan het aantekenschrift dat ongezien in het oud-papier verdween, dan het foldertje op de deurmat, dan het fotoalbum dat na twee generaties bij het vuilnis ging. Dat komt nu voor iedereen toegankelijk op internet. Dat wordt door Lyman en Varian allemaal als informatie meegeteld. Dat zorgt voor de bijna homeopathische verdunning van de voor ons echt waardevolle informatie. Dat is de echte informatie-inflatie.

[13] september 2001

Elsevier, Microsoft en de nieuwe economie

Toen de nieuwe economie nog bloeide, toen de dotcom-koersen elke belegger virtueel miljonair maakten en toen het begrip recessie voor eeuwig tot het verleden leek te behoren, toen haastten economen zich om verklaringen voor dit nieuwe fenomeen te bedenken. Hoewel ik nauwelijks meer van economie begrijp dan dat het eigenlijk onder de psychologie thuis hoort, schenen hun theorieën mij best redelijk te klinken. Anders dan in de oude economie, raken de basis-ingrediënten van de internet-economie, software en informatie, niet op als je er meer van maakt. En nog belangrijker: als je het eenmaal één keer hebt gemaakt, dan kost het nauwelijks meer om er nog 10, 1000 of 100.000 keer zo veel van te maken.

Nu de nieuwe economie even onvoorspelbaar en grillig blijkt te zijn als de oude, recessies nog altijd mogelijk blijken en beleggers niet slechts virtueel failliet gaan, nu hoor je niemand meer over die zo redelijk klinkende verklaringen. Toch zijn die net genoemde eigenschappen van de ingrediënten van onze informatiemaatschappij, die de economen als iets wonderbaarlijk nieuws ontdekt leken te hebben, nog altijd onverminderd waar. Alleen lag de causaliteit van oorzaken en gevolgen kennelijk toch wat anders dan die economen dachten.

Bij de grote wetenschappelijke uitgevers en bij Derk Haank van Elsevier blijkt van die theorieën toch wel iets te zijn blijven hangen. In NRC Handelsblad van 10 juli, gaf Haank een verklaring voor het besluit van een groep van zes uitgevers om de digitale versies van hun biomedische tijdschriften vrijwel gratis beschikbaar te stellen aan onderzoekers en bibliotheken in ontwikkelingslanden. Naast overwegingen van liefdadigheid - zelfs Kofi Annan kwam ter sprake - werd als belangrijkste argument genoemd, dat de overgang naar digitaal wel grote investeringen in apparatuur en software vergt, maar - inderdaad - "als dat eenmaal is terugverdiend, kan de laatste klant vrijwel zonder kosten worden aangesloten". Consequenties trekkend uit dat laatste argument, overweegt Elsevier inderdaad ook tijdschriften voor andere vakgebieden op deze manier aan te bieden.

Ik wil deze euforie van liefdadigheid niet meteen bederven met de flauwe vraag waarom ik dan niet die "laatste klant" kan zijn en die investeringen nu juist aan mij terugverdiend moeten worden. Want het is niet alleen in ontwikkelingslanden dat sommigen minder voor hetzelfde product blijken te betalen dan de grote gebruikers die al van oudsher duurbetalende klanten waren. En ik zal ook niet suggereren dan maar een proxy-server in een ontwikkelingsland neer te zetten om net zo goedkoop gebruik te kunnen maken van die Elsevier tijdschriften. In plaats daarvan blader ik verder in dezelfde aflevering van de NRC waarin Haank zijn verhaal mocht doen.

Wat een toeval: twee bladzijden verder, op de opiniepagina staat een vette kop "Microsoft moet iedereen gratis Windows leveren". Inderdaad: daarvoor geldt precies hetzelfde financiële argument als door Derk Haank geformuleerd werd. Alleen was het hier niet Bill Gates zelf, die een onverwachte vorm van liefdadigheid jegens zijn klanten bepleitte. Nee, nu de rechter heeft beslist dat Microsoft niet hoeft te worden opgesplitst, zag een Amerikaanse hoogleraar en oud-minister als enig alternatief voor het doorbreken van de monopoliepositie van dat bedrijf, dat het verplicht zou moeten worden Windows voortaan gratis beschikbaar te stellen. Windows kan dan een standaard besturingssysteem blijven, op dezelfde manier als meer dan een eeuw geleden de vorm van stopcontacten gestandaardiseerd is (denkt die Amerikaan, naar de platte pootjes van de stekker van zijn eigen schemerlamp kijkend).

Niet dat die Amerikaanse hoogleraar overigens zelf gelooft dat zijn scenario - onder de regering Bush - ooit werkelijkheid zal worden. Bij de uitgevers is dat wel al gebeurd en - ondanks een beetje vergelijkbare monopolieposities - zelfs geheel vrijwillig, zonder justitiële druk. Nu maar naar de beurskoersen kijken wie van beiden de echte nieuwe economie uiteindelijk het best begrepen blijkt te hebben, Bill of Derk.

5, 50, 500 en andere ronde getallen

In 1502, nu precies 500 jaar geleden vertrok Columbus voor zijn vierde en laatste reis naar Amerika. Het nieuwe continent was op dat moment al enige tijd ontdekt. Maar pas dit keer werden voor het eerst cacaobonen mee teruggenomen naar Europa en tegelijk begonnen Spanjaarden ook slaven te verschepen vanuit Afrika. In datzelfde jaar 1502 werkte Michelangelo hard aan beeldhouwwerken voor de dom van Siena en werd ook de geit geïntroduceerd op het eiland Mauritius. Allemaal redenen voor klinkende jubilea? Nou, niet echt, geloof ik eigenlijk. En hoe ik dit allemaal weet? Zoveel historisch culturele belangstelling had u nooit achter me gezocht? Och, gewoon "1502" als zoekterm in Google intikken en al dit soort wetenswaardigheden rollen online over je scherm.

Eind november 1979 ging men ook al online. Maar dergelijke trivia waren op dat moment nog niet zo makkelijk uit computers op te vragen. Omdat online toen nog een beetje langzaam ging met 300 bits per seconde via ruisende telefoonlijnen naar Lockheed en ESA. En omdat er nog niet veel meer te doorzoeken viel dan alleen een paar zeer serieuze wetenschappelijke bibliografische databeesjes. Maar er was toen wel voor het eerst een cursus om dat online zoeken te leren aan nieuwelingen in het zoekvak. Een cursus gegeven en georganiseerd door de voortrekkers van VOGIN, en vooral gerealiseerd dankzij het doorzettingsvermogen van Lieuwien Koster uit Wageningen. Het was op dat moment immers nog allerminst simpel om genoeg terminals, modems en telefoonlijnen bij elkaar te krijgen, om temidden van een spaghetti aan kabels, zo'n vijfdaagse cursus te kunnen organiseren in een congrescentrum waar zelfs nog nooit iemand van het woord "online" gehoord had.

Zelf was ik toen nog een eenvoudig natuurkundige. Maar vorige maand, 22 jaar later, was ik er wel bij toen diezelfde VOGIN-cursus voor de 50ste keer gegeven werd. Diezelfde cursus? Natuurlijk is de inhoud voortdurend met zijn tijd meegegaan, anders had hij het geen 50 keer volgehouden, maar hij duurt wel nog altijd vijf dagen en hij wordt nog altijd gegeven door ervaren vakgenoten. En het blijft evengoed opmerkelijk dat zo'n hoogtechnologische cursus over een periode van meer dan 20 jaar, met ijzeren regelmaat, vrijwel elke keer volgeboekt, al 50 keer met succes is gegeven. Reden voor een klinkend jubileum? Eigenlijk wel degelijk.

Januari 1997, nu precies vijf jaar geleden. Het eerste nummer van Informatie Professional viel in de bus. De redactie gaf zijn credo af in een stukje onder de titel "Van haarvaten en bypasses". Daarin werd geconstateerd dat bibliotheekwezen en documentatiesector convergeerden en dat traditionele online en Internet naar elkaar toegroeiden. Open en Login als twee afzonderlijke bladen leek te veel versnippering. De "bypasses" uit de titel verwezen naar de afsteekjes, dwarsverbindingen en rolwisselingen die in de informatieketen ontstonden.

Ook wees men op het ontstaan van steeds meer laagdrempelige verbindingen en diensten voor informatiegebruikers, de "haarvaten" van de informatievoorziening. Eigenlijk vinden we al die dingen intussen zo gewoon, dat het achteraf bijna verbazing wekt dat het toen nog zo expliciet gezegd moest worden. Net zo gewoon is intussen de benaming "informatie professional" geworden voor de informatiewerkers voor wie IP bedoeld was. Dat onze beroepsgroep ooit zonder die aanduiding door het leven kon, valt haast niet meer voor te stellen.

We lazen in dat eerste nummer al over personeelsmanagement in digitaal perspectief, over kennisloket en kennistransfer en over metawebwoelers. Ook werd gemeld dat NBBI integreerde met TNO-STB. Wat daarvan geworden is, weten we intussen. De oud-NBBI'ers van toen bevolken nu in groten getale de redactie en de redactieadviesraad van Informatie Professional, terwijl er haast geen een meer bij TNO te vinden is. Een volgend nummer meldde de - achteraf veel succesvollere - overstap van Maria Heijne van Surfnet naar TU Delft. Dutchess heette toen nog NBW, Nederlandse Basisclassificatie Web, en het inrichten van een website was een onderwerp waaraan een heel artikel gewijd kon worden. Voor wie aan onze huidige vormgeving en inhoud gewend is, zien die eerste nummers er achteraf gezien nog niet zo erg gelikt uit en misschien zelfs al een beetje achterhaald. Maar toch ook wel al een heel verschil met de beide ouders Open en Login.

Vijf jaar IP reden voor een klinkend jubileum? Wis en waarachtig wel. Maar daarmee vallen we onze lezers nu niet verder lastig. Voor die lezers moeten we eerst maar weer een goed volgend nummer maken en zorgen dat we met onze tijd blijven meegaan, net als die VOGIN-cursus. Zou het dan lukken 50 jaargangen te halen?

Over concurrentie en monopolies

Reed-Elsevier was vorige maand weer even hevig in het nieuws met de welles-nietes geruchten dat alsnog een samengaan met Wolters-Kluwer op handen was. Derk Haank, directeur van de wetenschappelijke uitgeefpoot van het concern, greep zijn aanwezigheid op een bibliotheekcongres, begin februari in Bielefeld aan, om nog eens duidelijk te maken dat het wat hem betreft in elk geval *nietes* was. Hij stelde onomwonden op dit moment geen enkele behoefte te hebben verder te groeien. Zelfs in overname van kleine wetenschappelijke uitgevers was hij niet geïnteresseerd, 1200 tijdschrifttitels was voor hem voorlopig genoeg. Eerst moesten zijn concurrenten maar eens wat verder groeien.

Zeker een geruststelling voor mijn baas, Bas Savenije, die voorlopig geen opgewonden journalisten meer aan de telefoon zal krijgen, die beslist willen weten "wat hij daarvan vindt". Moet je altijd ergens iets van vinden? De suggestie van onbetamelijke monopolievorming die in dergelijke vragen doorklinkt, berust natuurlijk op een onjuiste vergelijking met de broden van de bakker op de hoek. Wie een brood wil kopen kan gebruik maken van de concurrentie tussen de bakkers in de buurt. Op een wetenschappelijk tijdschrift heeft elke uitgever altijd een monopolie, of het nu een grote uitgever als Elsevier is of een kleintje met een enkele titel. Geen enkele andere uitgever heeft immers diezelfde titel. Zoals ook onze Otto het monopolie op IP heeft.

Zelfs de altijd wat serieuze Duitsers kreeg Derk Haank met zijn voordracht uit de plooi. Hij benadrukte nog eens - wat we als klanten al gemerkt hadden - dat hij best nieuwe business-modellen wil uitproberen. Maar té bont moet je dat natuurlijk ook niet maken. Aan een situatie waarin bibliotheken (geld) uitgeven en hij dat geld ontvangt, zijn we zo gewend dat hij dat maar liever zo wilde laten ;–). Wel maakte hij nog eens duidelijk dat aan hetzelfde product voor verschillende klanten best een verschillend prijskaartje kan hangen - ik meldde dat vorig jaar al eens in een column (september 2001).

Op diezelfde Bielefeld 2002 Conference werd trouwens ook duidelijk dat bibliotheken steeds meer geneigd zijn de concurrentie met het uniforme Google-interface aan te gaan. Alle leveranciers van metasearch portal software, waarmee een consortium van Nederlandse UKB-bibliotheken in onderhandeling (geweest) is - en zelfs nog een paar meer - waren in Bielefeld vertegenwoordigd. Met dergelijke systemen kun je meer catalogi en databases gelijktijdig doorzoeken. Iets waar we vijf jaar geleden in Utrecht al op een andere manier mee bezig waren, één interface en één zoekgang voor zoveel mogelijk bronnen, blijkt intussen door iedereen te worden nagestreefd.

Ook de meestal meteen daaraan gekoppelde *reference linking* wordt steeds algemener. Het vanuit een bibliografische database doorlinken naar de volledige tekst van een artikel of het doorlinken tussen full-text wetenschappelijke artikelen via hun literatuurreferenties. Via het OpenURL mechanisme bieden steeds meer software- en informatieleveranciers SFX-achtige diensten. Het Israëlische ExLibris lijkt, sneller dan ik had verwacht, zijn monopolie op dit terrein kwijt te raken. Zelfs online hosts kunnen niet meer achterblijven. In navolging van STN kondigde deze week ook Dialog aan dat ze *reference linking* vanuit haar databases aanbiedt.

Al deze concurrentie maakt het voor ons, grootschalige informatiegebruikers, nog altijd niet echt goedkoper, maar het draagt in elk geval wel bij aan een gestage verbetering van de functionaliteit van de zoek- en informatieproducten die wij onze klanten kunnen aanbieden.

Werkeloos

Al een flink aantal jaren geef ik cursussen hoe je op internet kunt zoeken. Ik vrees nu echter binnenkort, wat betreft die cursussen, werkeloos te worden. Recent is namelijk een boek verschenen waar alles al in staat. En dat *alles* bedoel ik dit keer beslist niet cynisch, want het is gewoon waar. Het boek is bovendien ook nog in het Nederlands geschreven. Collega Marten Hofstede heeft zijn niet geringe kennis en ervaring met zoeken op het web gebundeld in Het zoekboek voor het web. Eigenlijk heb ik niets kunnen bedenken dat in mijn cursussen aan de orde komt en dat niet ook in dit boek te vinden is. Vandaar die nakende werkeloosheid.

Overigens wel dapper om een statisch papieren boek te schrijven over zoiets vluchtigs en dynamisch als het web. Wat je vanmiddag schrijft is soms vanavond al weer achterhaald doordat een dienst verdwijnt, een interface verandert of een nuttige zoekfunctie er plotseling niet meer is. Dat NorthernLight in het boek nog als gewone zoekmachine wordt genoemd, zal ik Marten dus zeker niet aanrekenen. Bovendien is er wel nog een website aan dit boek gekoppeld waarop de auteur zal proberen ons up-to-date te houden. Dat is iets dat geen modern leerboek meer lijkt te kunnen missen, ook als het om minder vluchtige onderwerpen gaat.

Ik mag hier trouwens wel eens een voorbeeld aan nemen. Als eindredacteur van Online opsporen van informatie, het leerboek dat van oudsher het zoeken in professionele informatiebronnen behandelt, zit ik er namelijk al jaren tegenaan te hikken hoe in een nieuwe druk het geheel veranderde informatielandschap en de daarin beschikbare zoektechnieken systematisch, overzichtelijk en ook nog een beetje volledig kunnen worden weergegeven. Met daarbij als voornaamste vrees, dat wat ik vandaag opschrijf, zelfs voor die professionele bronnen, overmorgen al weer achterhaald kan zijn. En bovendien bevat het web zelf ook zoveel professioneel nuttige informatie, dat dat als informatiebron niet onbelicht mag blijven. Maar gelukkig kan ik voor details daarover nu makkelijk naar dat nieuwe boek verwijzen.

Toch zou het feit dat er nu wel een zoekboek voor het web is, maar nog geen vernieuwd zoekboek voor het zoeken in professionele databases, onbedoeld kunnen bijdragen aan het onjuiste idee dat *alles* - gratis - op het web te vinden is. Of erger nog, dat die dingen die je niet gewoon op het web kunt vinden, eigenlijk niet de moeite waard zijn gevonden te worden. Dat is ook één van de aspecten die ten grondslag liggen aan het thema van de komende IP-lezing. Iedereen kan heel makkelijk alles zelf op het web vinden. Google is zo slim dat je geen professionele zoekers en geen professionele gestructureerde en gecontroleerd ontsloten informatiebronnen meer nodig hebt. Dat we bij dat "*makkelijk alles op het web kunnen vinden*" overigens best een heel groot vraagteken mogen zetten, blijkt wel uit het feit dat Marten Hofstede meer dan 300 bladzijden vol kan schrijven met nuttige tips en webwetenswaardigheden waar zelfs een informatieprofessional nog genoeg van kan leren. (En uit het feit dat mijn zoek-cursussen - tot dusverre - goed bezocht werden).

Toch stellen we bij de IP-lezing en in het bijbehorende themanummer, oktober a.s., onder meer de vraag of de professionele zoeker wellicht werkeloos wordt, door betere informatievaardigheden van de klanten en gebruikers, door slimmere software en doordat steeds meer mensen zich met een paar weetjes van het web tevreden lijken te stellen. En even gewettigd is dan de vraag of de aanbieders van professionele informatie - databaseproducenten, hosts, uitgevers - om die zelfde redenen niet eveneens werkeloos dreigen te raken (al noem je dat in de zakenwereld anders).

Zelf ben ik voorlopig zeker niet werkeloos want dat nieuwe zoekboek voor het vinden van informatie in professionele - betaalde - informatiebronnen moet er nu toch echt heel snel komen.

Zoeken uitgezocht

VOGIN, de vereniging van de echte zoekers, bestaat dit jaar 25 jaar. Zo lang - en zelfs al wat langer - wordt in Nederland dus al professioneel online naar informatie gezocht. Ter gelegenheid van dat VOGIN-jubileum is een CD-ROM uitgekomen met daarop de volledige teksten van alle jaargangen van LOGIN, het lijfblad van de VOGIN-leden. Dat waren "maar" 20 jaargangen, want LOGIN was een van de twee bladen waaruit vijf jaar geleden het huidige Informatie Professional is voortgekomen.

Op die CD-ROM kwam ik ook een oud artikeltje van mezelf tegen, het eerste in een serie over zoekmachines, een serie die de voorloper van onze huidige www-rubriek was. Hoewel oud? 1995 is in feite nog niet zo heel lang geleden. In dat artikel rapporteerde ik opgetogen over mijn eerste ervaringen met Lycos, een pas door mij ontdekte zoekmachine voor het web, waarin maar liefst 1,5 miljoen webpagina's doorzoekbaar waren gemaakt. Toch was Lycos ook op dat moment al niet de enige zoekmachine - al was het wel verreweg de grootste.

Een paar andere, op dat moment in mijn ogen interessante zoekproducten waren onder meer WebCrawler, WWWWorm, W5, CUI-W3, Jumpstation en NIKOS.
Kent u die al? ..... Nee?
Dat is dan niet zo verwonderlijk, want van al die namen is er, met uitzondering van Lycos en WebCrawler, zeven jaar later niet één meer over. En Lycos en WebCrawler zijn alleen nog namen, want tegenwoordig maakt Lycos voor zijn zoekfunctie gebruik van de index van FAST-All-the-web en is WebCrawler intussen omgevormd tot een metacrawler die zijn resultaten bij (eveneens) FAST, Inktomi, AskJeeves, LookSmart en nog wat andere zoekdiensten vandaan haalt.

Zelf ben ik intussen wel al een beetje oud en 1980 is wel al lang geleden. In dat jaar kwam ik in het vak en bij VOGIN terecht. Toen leerde ik online zoeken. Een van de eerste zoeksystemen die ik ging gebruiken was Dialog - door de echte harde kern toen meestal nog "Lockheed" genoemd, naar het vliegtuigconcern waar dit zoeksysteem ontwikkeld was. Met een 300 baud modem via inbelpunten van Dabas, een datacommunicatiedienst van PTT, kon je verbinding maken met Amerika. Na ingewikkelde reeksen inlog-procedures kon je commandogestuurde zoekacties doen via een domme terminal.

Op dat terrein moet in die 22 jaar dus helemaal wel veel veranderd zijn. Inderdaad, bij alle online hosts kun je intussen ook via web-interfaces zoeken. Maar als ik echt even snel, heel gericht en systematisch zoeken wil, dan kan ik nog altijd via Telnet mijn Dialog-zoekacties doen. Met nog vrijwel exact dezelfde commando's als 22 jaar geleden, ook al is Dialog sindsdien al drie of vier keer aan andere concerns doorverkocht. Er zijn hooguit drie of vier echt nieuwe commando's bijgekomen, sommige zijn wat gestroomlijnd en je kunt intussen niet meer in 60 maar in 600 bestanden zoeken. Heerlijk die stabiele online hosts, waar niet elke paar jaar alle oude zoekmogelijkheden verdwenen blijken te zijn, zoals op het web. Ideaal voor de echte zoekers. *Old searchers never die.*

Spookrijden

Het Amsterdamse Amstelstation wordt al een hele tijd verbouwd en gerenoveerd. Voor het gemak was vorige week de gewone trap naar een van de perrons plotseling afgesloten. Wel bleken aan de andere kant intussen twee splinternieuwe roltrappen te zijn verschenen. Alleen deden die het nog niet. Nu is lopen over stilstaande roltrappen een van de vervelendste dingen die ik ken, met aan begin en eind verhoogde struikelkans door die langzaam verlopende treehoogtes. Bovendien manoeuvreerde in de drukte ook nog iemand met een fiets onhandig op de schouder net voor me de trap op, terwijl de trein elk moment kon binnenkomen. En die wilde ik beslist niet missen. Gelukkig was de linker roltrap er ook nog. Flink doorstappend, maar ook een beetje in gedachten verzonken omhoog. Daardoor duurde het even voor ik merkte dat dat gezeul met de fiets op de trap naast mij, eigenlijk net zo hard opschoot als ik. Inderdaad, u begrijpt het al, de andere roltrap was alsnog in beweging gekomen. Ik had helemaal niet zo'n haast hoeven maken.

Het duurde daarna nog minstens tien treden, voor ik door had dat ik zelfs in absolute zin lang niet zo hard opschoot als op grond van mijn klimtempo te verwachten viel. Slimme lezers hadden natuurlijk meteen al begrepen dat mijn roltrap mij als spookrijder had herkend en zich ook al lang in beweging had gezet, maar dan wel de andere kant op. Een ultieme tempoversnelling bracht me toch nog boven. De paar laatste opstapjes werden me daarbij noodlottig, zodat ik al struikelend plat voorover op het liftbordes belandde, temidden van de besmuikt toekijkende meute op het perron. Deze malloot vormde een welkome afwisseling bij het wachten op hun vertraagde trein.

Steeds meer vakgenoten lijken zich net zo te voelen als ik vorige week op die roltrap. Je doet voortdurend flink je best om bij te blijven met alle informatie en nieuwe ontwikkelingen op ons vakgebied. Maar hoe hard je ook loopt, je schiet niet echt op, alsof je voortdurend tegen de stroom in moet. En als je even verslapt, ben je weer helemaal beneden aangeland.

Gelukkig zijn er leuke bladen die ons wel een beetje helpen om bij te blijven, zoals bijvoorbeeld Informatie Professional. Ook zijn er af en toe collega's die je via discussielijsten op de hoogte houden, zoals bijvoorbeeld via Nedbib-L. Daar wees collega-redacteur en collega-docent Jan Verhoeven recent op het interessante oktober-nummer van het digitale tijdschrift Information Research. Een nummer dat "intentionally provocative" bedoeld was en eens met een kritische blik naar kennismanagement keek.

De eindredacteur van het blad vond dat begrip maar een "conceptually empty buzz word". Andere auteurs zagen daarin alleen een nieuw labeltje voor het aloude informatie management, omdat je kennis niet kunt vangen, codificeren of vastleggen. Een analyse van een representatieve doorsnee van de literatuur en van websites van consultancy bedrijven en MBA opleidingen, leverde weinig steun op voor het idee dat er ook maar ergens iemand bezig is met iets dat je werkelijk het managen van "kennis" mag noemen. Andere auteurs hadden nog wel behoefte aan wat nader onderzoek om te bepalen of kennismanagement wellicht toch wat meer is dan een "broad-shouldered fad". Daar had ik wel even Van Dale bij nodig om bevestigd te krijgen dat een "fad" een gril of een stokpaardje is. Maar breedgeschouderd?

Dit alles sluit aardig aan bij een eveneens recente column in Computable, waarin aanbieders van softwarepakketten werd aangeraden hun producten absoluut niet langer te promoten als krachtige hulpmiddelen voor "knowledge management". Dat begrip is helemaal uit. "Content management" is intussen HET nieuwe buzz word, waarmee je je producten aan de man moet brengen. Het is maar dat u ook dat vast weet.

Je mag je achteraf dus afvragen of het wel zo handig was van IDM-opleidingen om zich in hun gezamenlijke missie-notitie "Focus op kennis", zo volledig en zonder voorbehoud alleen nog maar achter het label "kennis" op te stellen. Soms is het maar beter niet te proberen tegen roltrappen in omhoog te lopen. Soms kun je je maar beter stilletjes weer naar beneden laten terugvoeren en rustig met al die andere mensen mee omhoog gaan. Alleen is het zo jammer dat je nooit te voren weet wanneer dat beter is, wanneer je met je haast alleen maar kans loopt om boven aangekomen plat op je gezicht te gaan en wanneer je door af te wachten juist de trein zult missen. Alleen bij stationsroltrappen, daar weet ik het nu wel. De trein heeft toch altijd vertraging.

Domesday

In de jaren tachtig van de vorige eeuw kwam ik bijna jaarlijks op de Online Conferentie in Londen. Zo ook in 1986. Dat jaar kwam ik terug met enthousiaste verhalen over een schitterend project van de BBC, dat op de bijbehorende beurs te zien was - als je althans kans zag er een glimp van op te vangen, over de schouders van al die andere enthousiaste bezoekers. Ter gelegenheid van de 900ste verjaardag van het "Domesday Book" had de BBC een veelomvattend project opgezet om met de modernste technologie een moderne Domesday-versie samen te stellen.

Het oude Domesday Book was een beschrijving van de 13.418 nederzettingen in het Engeland van het jaar 1086, samengesteld in opdracht van Willem de Veroveraar - er is uiteraard ook een website (www.domesdaybook.co.uk). De inhoud van het nieuwe Domesday Book stond op twee 12 inch interactieve videodisks - 30 cm, zo groot als vinyl LP's! Om het te kunnen raadplegen moest een videodisk-speler worden aangesloten op de BBC Acorn microcomputer die toen door de BBC werd gepromoot om ook het grote publiek aan de microcomputer te krijgen. Volgens de (moderne) overlevering zou je zeven jaar nodig hebben, als je alles zou willen bekijken: 250.000 plaatsnamen, 25.000 kaarten, 50.000 foto's, 3000 datasets met statistische en andere gegevens, en één vol uur bewegende video. (Zie bijvoorbeeld op www.atsf.co.uk/dottext/domesday.html).

Door de prijs van £ 2500 voor het geheel van schijven plus apparatuur heeft dat moderne Domesday Book nooit een echt grote verspreiding gehad, zeker in Nederland niet. Het is vooral terecht gekomen op Engelse scholen en in Engelse bibliotheken. Ondanks mijn enthousiasme van destijds, had ik er dan ook nooit meer aan terug gedacht. Totdat afgelopen december ineens weer berichten over het Domesday Project in de krant verschenen. Als in een soort archeologisch project waren informatici erin geslaagd om alle data én de bijbehorende software weer op te graven en te reconstrueren, nadat die bedolven waren geraakt in niet langer gebruikte data-formaten, op een niet langer courant medium, voor niet langer werkende videodisk-drives die aan niet langer gebruikte computerhardware gekoppeld moest worden.

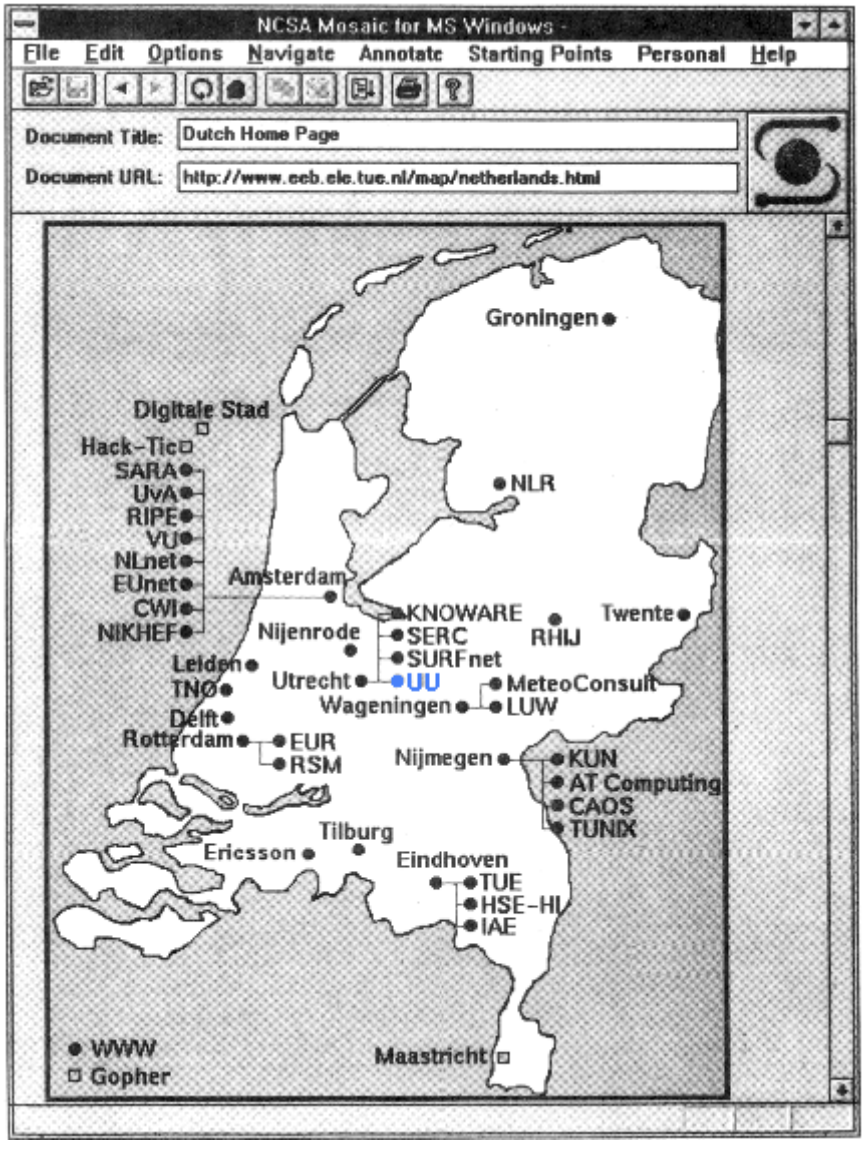
De geschiedenis van dit Domesday Book is een wel heel aansprekend voorbeeld van het probleem van digitale duurzaamheid. Dat moderne Domesday Book, waarvan toch vele duizenden exemplaren geproduceerd zullen zijn, was na 15 jaar in feite al weer helemaal verdwenen, want "onleesbaar" geworden.

Terwijl het oude Domesday Book, waarvan de inhoud nog dateert van ver voor de uitvinding van de boekdrukkunst, na 900 jaar nog altijd bestaat en "leesbaar" is.

Toch is dit voorbeeld volgens mij niet echt representatief voor de grote meerderheid aan digitaal opgeslagen informatie. Die eerste experimenten en projecten met digitale media en interactieve computersystemen waren natuurlijk ontzettend interessant. Maar de uit dat tijdperk afkomstige hoeveelheid digitale informatie vormt maar een minieme fractie van wat er intussen aan digitale informatie geproduceerd is. Die kleine fractie is echter ontwikkeld voor een veelheid aan heel specifieke systemen, met maar weinig toepassingen voor elk van die systemen. Daar ligt dus een probleem: het loont commercieel niet erg om die informatie (softwarematig) leesbaar te houden.

Bij de grote meerderheid van de digitale informatie hebben we echter te maken met de wetten van de grote aantallen. De miljoenen - of zelfs wel miljarden - documenten in allerlei WordPerfect- en Word-versies vormen een zo grote potentiële klantenbasis voor allerlei softwarebedrijven, en er zijn zoveel organisaties die groot bedrijfsmatig belang hebben bij blijvende toegankelijkheid van die documenten, dat daarvoor wel readers, viewers en achterwaarts compatibele systemen beschikbaar zullen blijven. Dat neemt overigens niet weg dat we natuurlijk voortdurend op ons hoede moeten blijven - en zeker wat betreft de opslagmedia - dat geen informatie verloren gaat. Maar zo extreem als met die eerste systemen zal dat zich wel niet meer voordoen.

Een heel ander probleem - dat in het kader van de digitale duurzaamheid intussen gelukkig ook wordt onderkend - vormen al die informatiebronnen waarvan de "content" voortdurend wordt overschreven door ge-update nieuwere versies. Al die webpagina's die voortdurend worden vernieuwd. Wie weet nu nog hoe de "Dutch Home Page" er ooit uitzag? Een webpagina met een kaart van Nederland, waarop direct aanklikbaar alle 40 websites van Nederland waren aangegeven. Hét beginpunt van iedere websurftocht begin jaren negentig. Ik heb hem nog! Nee niet als HTML-file. Gewoon een afbeelding van die webpagina uit de papieren krant, september 1994, die ik had uitgeknipt. Zeer duurzaam. Kijk maar. Tot aan de dag des oordeels ligt die toestand daarmee vast.



Ontploffende komkommers

Vlak voor mijn vakantie leek zelfs op de Nedbib-lijst de komkommertijd al aangebroken. Uitgebreide discussies over in PC's exploderende CD's wekten de indruk dat niemand serieuzer werk te doen had. Toch bleek dit door mederedacteur Adriaan van Geest aangesneden probleem - uiteraard - heel serieus. Het gebeurt namelijk echt, en niet alleen bij de in Adriaan's bibliotheek geleende schijfjes. Iedereen kwam met eigen ervaringen of met op internet gevonden verhalen. En daar waren hele leuke bij. Vooral voor een voormalig experimenteel fysicus zoals ik.

Het probleem komt voort uit de steeds grotere leessnelheid van CD-drives. De oorspronkelijke audio-CD was uiteraard bedoeld om altijd op dezelfde snelheid te worden afgelezen. Anders dan bij het aloude vinyl, betekent dat overigens niet dat het schijfje steeds even snel ronddraait. Het aantal bits per afgelezen cm putjesspoor is bij CD's namelijk over de hele schijf constant. Aan de buitenkant van de schijf is de lengte van een hele omwenteling echter bijna drie keer zo lang als vlakbij het gaatje in het midden. Om daar even snel te kunnen lezen als aan de buitenkant, moet de schijf dus bijna drie keer zo hard ronddraaien. Gewone audio-CD's hebben daarom een snelheid die varieert tussen 200 en 530 omwentelingen per minuut. Voor moderne computertoepassingen is de corresponderende leessnelheid van ruim 170 kB/sec echter lang niet genoeg. Vandaar dat CD-drives in PC's steeds sneller zijn gaan draaien; de nieuwste zelfs al 64x zo snel als audio-spelers. Maar dat betekent dat schijfjes daarin ronddraaien met snelheden tot 34.000 omwentelingen per minuut, ofwel ruim 500 rondjes per seconde!

Daarom had men experimenten gedaan, waarbij schijfjes gemonteerd waren op een Dremel, zo'n handig klein boor-/slijp-machientje, of zelfs op een zware, met veiligheidskappen afgeschermde elektromotor. Experimentele testresultaten werden getoond in de vorm van foto's van tot scherven uiteengespatte CD's of van diepe zaagsneden die door weggeschoten CD's in het plafond waren achtergelaten, alsof een leger dol geworden cirkelzagen was ontsnapt. Om het wat wetenschappelijker te maken werden soms ook nog wat grafieken en staafdiagrammen toegevoegd.

Toch dreigde op de Nedbib-lijst even scepsis, toen één van de discussianten ontdekte dat de auteur van het meest wetenschappelijk uitziende verhaal daaraan de <title>-tag "Jorgen Stadje, Nonsens Page ..." had meegegeven. Moet je een web-pagina met een dergelijke titel serieus nemen? Gelukkig kun je terugvallen op simpele, maar degelijke middelbare school natuurkunde en zelf het destructief vermogen van dergelijke schijfjes narekenen. Weet u nog wel, de kinetische energie: $E=\frac{1}{2}mv^2$ (E is halfemveekwadraat). Quantummechanica en relativiteitstheorie zijn hier gelukkig niet voor nodig.

In een op maximum snelheid ronddraaiend schijfje in zo'n 64x drive, blijkt dan ongeveer 140 Joule aan bewegingsenergie te zitten. Dat is ruim anderhalf keer zo veel als in de alles splijtende service van Martin Verkerk. Die tennisbal is weliswaar vijf keer zo zwaar als een CD, maar de scherpe buitenrand van die CD, in dat krakkemikkige plastic laadje in de PC, haalt een snelheid van 770 km/uur. Daar kan Martin's service niet aan tippen. Nog een factor anderhalf sneller en CD's doorbreken de geluidsbarrière.

Nou gaat er natuurlijk niets mis, zolang dat schijfje maar keurig stabiel om zijn as blijft draaien. Maar als het gaatje niet heel precies in het midden zit of als er scheurtjes in het materiaal zitten, dan komen er enorme krachten op te staan. Ook middelbare school natuurkunde. Als je een CD in tweeën breekt en dan weer lijmt, dan moet die lijm een maximale trekkracht van 25 miljoen Newton/m^2 kunnen weerstaan, een trekkracht die in een gave CD intern door het plastic zelf wordt opgevangen. Nog wat aanschouwelijker geformuleerd: er zou een gewicht van 300 kilo aan die gelijmde breuk moeten kunnen hangen. Anders vliegen de twee helften - volgens Galilei - bij het draaien uit elkaar. De maan blijft ook alleen maar om de aarde draaien omdat de zwaartekracht er steeds aan trekt.

Schijfjes komkommer in CD-spelers leveren geen gevaar op; dat werkt alleen als sapcentrifuge. Met CD's erin worden de modernste spelers echter knap gevaarlijk. Al exploderen de schijfjes dan wel niet letterlijk, scherpe brokstukken blijken wel degelijk door het laadje heen naar buiten te kunnen ontsnappen - om van de schade aan het apparaat nog maar te zwijgen. En daar was het bij de klant van Adriaan gelukkig toe beperkt gebleven.

* PowerLabs High Speed CD-Rom Experiments - www.powerlabs.org/cdexplode.htm

PLoP, PLoM of PLoS

Auteursrecht blijkt weerbarstige materie, zeker in ons digitale tijdperk. En juristen zien altijd kans zulke problematiek nog weer ingewikkelder te maken dan het toch al is. Dat blijkt wel uit de discussie die de uitspraak van de Hoge Raad over KaZaA afgelopen december opriep. Gelukkig ben ik geen jurist, zodat ik ongestraft simplistische ideeën over dit soort onderwerpen kan ventileren. Aan de inhoud van deze column vallen door de lezer dan ook geen rechten te ontlenen.

In het kader van wetenschappelijke informatievoorziening brak Leo Waaijers onlangs in Informatie Professional (7/8-2003) al een lans voor nieuwe methoden van publiceren. Daarbij blijft het auteursrecht in elk geval bij de auteur berusten en fungeren diens artikelen niet langer uitsluitend als melkkoe voor de grote uitgevers. Door uitgevers opgelegde beperkingen op verspreiding van wetenschappelijke informatie hebben ook maar weinig met het feitelijke auteursrecht te maken. Dergelijke nieuwe modellen voor wetenschappelijk publiceren via Open Access tijdschriften maken de hele cyclus van de wetenschappelijke informatievoorziening goedkoper en zorgen tegelijk dat die informatie vrijelijk toegankelijk komt. Dat spoort prima met de al eerder vanuit de wetenschap zelf geformuleerde wens om te komen tot een PLoS, een Public Library of Science, waarin auteurs hun eigen publicaties vrijelijk beschikbaar stellen.

Je zou kunnen denken dat Digital Rights Management, een infrastructuur en software voor beheer van de rechten op digitaal beschikbare informatie, op dit terrein ook uitkomst zou kunnen bieden. In een waarschuwend stuk in NRC-Handelsblad, eind november, maakte Nol Verhagen (UB Amsterdam) zijn lezers echter snel duidelijk dat niets minder waar is. Alle uitzonderingsrechten die bibliotheken voor niet-digitale informatie van oudsher zijn toegestaan, kunnen door DRM en bijbehorende wetgeving wel heel makkelijk ongedaan gemaakt worden. En dat geldt dan voor alle soorten informatie waarmee bibliotheken te maken hebben.

Die andere informatie en media vormen mijn bruggetje naar de KaZaA-zaak, waarin de Hoge Raad uitsprak dat KaZaA zelf niet verantwoordelijk is voor het feit dat die software (ook) gebruikt wordt om er illegaal muziek mee uit te wisselen. Christiaan Alberdingk Thijm, onder meer advocaat van KaZaA, riep triomfantelijk dat dit "misschien wel de belangrijkste uitspraak op het gebied van internetrecht" was.

Dat professor Bernt Hugenholtz daarop reageerde met de bewering dat de voor KaZaA gunstige uitspraak alleen maar een gevolg was van knullig procederen door tegenstander Buma/Stemra vond ik niet zo interessant. Voor mij blijft dat "illegaal uitwisselen" namelijk het moeilijke punt. Ik zie nooit zo'n principieel verschil tussen het op cassette opnemen van een muzieknummer dat via de radio werd uitgezonden en het voor eigen gebruik downloaden van datzelfde nummer, ergens van internet.

Anderzijds ben ik ook wel weer zo genuanceerd dat ik onmiddellijk wil erkennen dat artiesten ook voor hun arbeid betaald moeten worden. Tot dusverre had ik ook altijd gedacht dat de problematiek van auteursrecht in de muziekwereld heel anders lag dan bij wetenschappelijk publiceren. Dat auteursrecht op platen en cd's zorgde dat artiesten hun terechte inkomsten kregen. De januari M-bijlage van NRC leerde me echter dat zoiets alleen maar het geval is als je Mick Jagger, Michael Jackson of Madonna heet. Onbekendere artiesten blijken meestal bitter weinig aan hun platencontracten over te houden en meestal helemaal niet meer ontvangen als de verkoop verdubbelt. Zogenaamd uit naam van de artiesten doen platenmaatschappijen dus vooral zielig ten behoeve van de eigen winst. Vandaar ook dat artiesten zelf hun muziek op internet beschikbaar beginnen te stellen, buiten die platenmaatschappijen om. Deels gratis te downloaden, deels tegen betaling en vaak ook om rechtstreekse verkoop van hun CD's te promoten, maar dan in elk geval wel direct ten bate van hun eigen portemonnee. Ondanks bepaalde verschillen met de PLoS, ontstaat er zo dus ook een soort van PLoM, een Public Library of Music.

En wat is dan de PLoP uit mijn titel? Acronymfinder.com geeft vier mogelijke betekenissen, maar geen daarvan is waar ik aan dacht. Een Public Library of Pleasure lijkt me niet zo'n goed idee, want internet biedt al meer dan genoeg *pleasure*, vooral voor een bepaald soort manlijke internetgebruikers. Plaatjeszoekmachines zorgen echter wel al heel lang voor een Public Library of Pictures. Dat we in geval van hergebruik van plaatjes ook wel degelijk met auteursrechten te maken hebben, laat onverlet dat eindeloze hoeveelheden plaatjes gratis te vinden zijn, én te bekijken zonder kans op juridische complicaties. Dat is door juristen gelukkig nog niet ingewikkelder gemaakt dan het was.

De Googlificatie van de hele wereld

Ik ben nog maar net terug van een rondreis door Florida. Daarbij heb ik zowel Disney World weten te vermijden, als - minder doelbewust - het gebied rond St. Petersburg en Tampa Beach. Ja, een St. Petersburg hebben ze in Florida ook al. Net als een Bagdad trouwens. Die laatste omtrekkende beweging was misschien toch niet zo verstandig. Niet alleen omdat achteraf bleek dat St. Petersburg een museum met een belangrijke Dali-collectie herbergt, maar ook omdat ik, thuis gekomen, nu door mijn directeur Peter van Gorsel geattendeerd moest worden op een fraai Flash filmpje van de Special Projects Division van het Museum of Media History in Tampa Bay Federal District.

In tegenstelling tot het Dali-museum, blijkt dit media-museum trouwens helemaal niet in het echt te bestaan. Het acht minuten durende Flash filmpje, te vinden op http://www.broom.org/epic/, kijkt terug vanuit het jaar 2014, het jaar waarin de New York Times offline ging. Met een sonore documentaire vertelstem verhaalt het filmpje van Robin Sloan en Matt Thompson over de historie van het ontstaan van EPIC, de 'Evolving Personalized Information Construct', een nieuw gepersonaliseerd medium, waarmee zoekmachine-gigant Google de rol van de klassieke nieuwsmedia heeft overgenomen:

> *Everyone participates to create a living, breathing mediascape. However, the Press, as you know it, has ceased to exist. The Fourth Estate's fortunes have waned. 20th Century news organizations are an after-thought, a lonely remnant of a not too distant past.....*
> *On Sunday, March 9 2014, Googlezon unleashes EPIC.*
> *The 'Evolving Personalized Information Construct' is the system by which our sprawling, chaotic mediascape is filtered, ordered and delivered. Everyone contributes now - from blog entries, to phone-cam images, to video reports, to full investigations.....*
> *EPIC produces a custom contents package for each user, using his choices, his consumption habits, his interests, his demographics, his social network - to shape the product.*

Het verhaal borduurt verder op de nu al bestaande belangstelling van zoekmachine-giganten - vooral Google - voor andere media en technieken, zoals kan worden afgeleid uit hun overnamebeleid en samenwerkingsverbanden. Belangstelling voor blogs en voor persoonlijke e-mail en sociale netwerken zoals Friendster. Belangstelling voor TiVo, de persoonlijke digitale bewaardienst voor televisie-programma's, en voor de technieken die Amazon gebruikt om, op basis van gedetailleerde kennis van de klant, aan diens specifieke smaak en belangstelling aangepaste aanbiedingen te kunnen doen.

Combinatie van deze trends zou uiteindelijk leiden tot de fusie van Google en Amazon tot Googlezon en de ontwikkeling van nieuwe mediaproducten. Dit alles gebaseerd op het ongeëvenaarde nieuwe Google Grid zoeksysteem, waarin iedere gebruiker ook zijn persoonlijke gegevens en bestanden kan opslaan. Dit alles resulterend in nieuwsberichten die dynamisch gegenereerd worden, voor iedere gebruiker persoonlijk, op basis van diens eigen voorkeuren, belangstelling, kennissenkring en smaak. Daar konden klassieke kranten uiteindelijk niet meer tegenop.

> *Today, in 2014, The New York Times has gone offline in feeble protest to Googlezon's hegemony, but Times has become a print-only newsletter for the elite and the elderly.*

Sommigen lijkt dit misschien de ultieme uitwerking van geautomatiseerde persoonlijke attenderingsdiensten, van de klassieke SDI's uit de jaren tachtig van de vorige eeuw. Voor anderen misschien juist de ultieme nachtmerrie, het vooruitzicht van nieuws- en informatiesystemen die hun gebruikers alleen nog maar naar de mond praten, die onder het mom van nieuwsvoorziening iedere gebruiker alleen nog vertellen wat die graag horen wil.

Ver doorgeschoten toekomstmuziek? Nu al kunnen we selectief alleen nog die blogs lezen - of op basis van zelf ingestelde filters "te lezen krijgen" - die afkomstig zijn van gelijkdenkenden elders op internet. Het daardoor niet meer openstaan voor opvattingen die afwijken van de eigen ideeën werd door Sunstein, in een aan blogs gewijd themanummer van de Communications of the ACM (vol. 47, no. 12, p. 57 - December 2004), zelfs al een gevaar voor onze democratie genoemd. Toch ook nog iets positiefs? In dit toekomstscenario heeft Google uiteindelijk wel Microsoft verslagen.

Overigens maar goed dat ik pas thuis over EPIC te horen kreeg. In de virtuele omgeving van mijn eigen PC was dat filmpje heel plezierig te bekijken en was ook de uitgeschreven commentaartekst op te zoeken - inderdaad met Google. In mijn tentje, tussen de armadillo's en de zeer reële muggen van Florida's kustmoerassen, was ik nog niet online (en had ik ook andere prioriteiten).

PS: Een volledig transcript van de gesproken tekst bij het filmpje is te vinden op masternewmedia.org

[23] februari 2006

Web 2.0 : de nieuwe kleren van het web

Als we kranten en weblogs mogen geloven, is recent een nieuwe versie van het web uitgekomen, release *Web 2.0*. Mogen we dan binnenkort Web 2.1 en Web 2.3-beta verwachten? Nee natuurlijk. Het is een wat kromme metafoor waarin handige marketing-jongens en *trend-watchers* het organische geheel van het web proberen te vergelijken met een softwareapplicatie waarvan met regelmaat een nieuwe release uitkomt.

Is er dan helemaal niets aan de hand? Zeker wel. Al vijftien jaar is er voortdurend wat aan de hand met het web. Zoals uiterlijk en verdere techniek van auto's de eerste twintig jaar van hun bestaan sterker zijn veranderd dan ooit daarna, zo is dat met het web ook het geval. En dat gaat met golven. Een van die laatste golven zou je die van de doe-het-zelvers kunnen noemen. Hierbij mijn doe-het-zelf-top-tien.

1. Iedereen nu zijn eigen journalist. Het journalistieke monopolie is doorbroken door de weblogs van iedereen die denkt wat te melden te hebben. Nog afgezien van het punt of die gedachte juist is, prangt wel de vraag wie al die blogs uiteindelijk moet lezen, tegen de tijd dat iedereen zijn dagboek aan internet toevertrouwt. Ook daarvoor zijn al oplossingen. Er zijn gespecialiseerde weblog-zoekers als *Technorati* en ook *Google-blog search*, om selectief alleen dat te selecteren dat je bevalt. En anderzijds *Digg.com* als centrale plek waar iedereen zijn nieuwsberichten kan plaatsen en waarin populariteit op zijn Googliaans de presentatievolgorde bepaalt. Of *Newsvine*, waar officieel nieuws met "amateur-nieuws" gemengd kan worden.
2. Iedereen ook zijn eigen fotograaf. Niet alleen dat we met digitale camera en mobieltje makkelijker dan ooit miljoenen foto's kunnen maken. Met websites als *Flickr.com* kunnen we die ook eenvoudig publiek maken. Zelfs reguliere kranten blijken er af en toe al dankbaar gebruik van te maken.
3. Iedereen zijn eigen bibliothecaris. Met *Furl* kun je elke op het web gevonden informatie eenvoudig opslaan en terugvindbaar maken. Ook geen van boven opgelegde taxonomieën of thesauri om informatie te ontsluiten, maar ieders persoonlijke tagging die ook door anderen gebruikt kan worden. Social bookmarking of folksonomies voor het ontsluiten van al die foto's op *Flickr.com* of van de berichtjes in *Digg.com* en *Newsvine* of van verzamelde bookmarks op *del.icio.us*.
4. Iedereen ook zijn eigen encyclopedist. We kunnen allemaal voor Voltaire of Diderot spelen door zelf lemma's aan de *Wikipedia* toe te voegen over onderwerpen waarvan we verstand denken te hebben - of door de fouten in andermans teksten te verbeteren.
5. Iedereen zijn eigen boek- of filmrecensent. Boekrecensies bij *Amazon* kenden we allemaal al; bij *Movies.yahoo* worden officiële filmrecensies nu ook aangevuld met recensies door willekeurig wie.
6. Iedereen zijn eigen (sociale) netwerkbouwer met systemen als *Hyves* of *Orkut*. Of denk aan *MySpace*, dat vooral populair is bij jongeren om daarin hun eigen - desnoods fictieve - web-identiteit op te bouwen.
7. Iedereen ook zijn eigen veilingmeester op *e-Bay* of *Marktplaats.nl*. Je blijkt daar zelfs succesvol te kunnen voorwenden je eigen kind te willen veilen.

8. Iedereen zijn eigen radio- of tv-station. De Tsunami van vorig jaar was al de meest uitgebreid gedocumenteerde ramp ooit, dankzij de video's van doe-het-zelvers. Via *Google-video*, *Blinkx.tv*, *Singingfish* en soortgelijke diensten kan iedereen zijn persoonlijke video's op het web aanbieden. En daarnaast natuurlijk *Podcasts*, al dan niet gekoppeld aan weblogs en *RSS-feeds*, als medium om je audio en video meer actief te verspreiden.

9. Iedereen ook zijn eigen zoekmachine. Met *Rollyo* geef je de betrouwbare domeinnamen op waartoe je je zoekacties wilt beperken. En die eigen "searchrolls" kun je natuurlijk ook weer heel sociaal met anderen delen.

10. En bijna iedereen zijn eigen software. Web-services en lichtgewicht mini-applicaties of *widgets* die naar je desktop of je browser komen via technieken als *AJAX* (Asynchronous JavaScript And XML). *Google-maps* werkt daarmee en biedt anderen zo de mogelijkheid daar hun eigen geografische dienstverlening bovenop te zetten. Ook andere interactieve diensten en zelfs web-based tekstverwerking zijn op dit soort technieken gebaseerd.

Hoewel we zo nog wel even door hadden kunnen gaan, zal er nooit concreet iets als Web 2.0 bestaan. Maar al deze voorbeelden getuigen wel van nieuw elan, van een opnieuw aangezwollen golf aan nieuwe ideeën, bedrijven en technieken, nu de ergste schrik van het barsten van de dot.com bel, al weer ruim vier jaar geleden, is weggeëbd.

De combinatie van doe-het-zelf, personalisatie en collaboratie - maar dan in de positieve betekenis van "samenwerking" - daar draait dit allemaal om. In elk geval zit het web dus weer stevig in de nieuwe kleren, zo stevig zelfs dat de meeste van de hier genoemde diensten intussen zijn overgenomen door de grote spelers in het wereldwijde mediagebeuren, door Yahoo of Rupert Murdoch, door Google of MSN. En zelfs AJAX wint zo toch weer een keer.

PS: nog maar weinig van de hier genoemde diensten bestaan anno 2018 nog.

De privacy van aardappels afgieten

Vorige maand publiceerde Google een lijstje meest in zijn index voorkomende woorden. Niet verbazend dat "a" en "the" op 1 en 2 stonden. Elk kwam voor in 25 miljard webpagina's. Wat wel verraste waren "copyright" en "privacy", als vrijwel eerste niet-stopwoorden op 27 en 29, in 13 miljard webpagina's. Komen "a" en "the" echt maar twee keer zo vaak voor als "copyright" en "privacy"?

Navrant, die twee woorden markeren precies de omslag waarmee Google, publiekslieveling af, in de beklaagdenbank terecht gekomen lijkt. Voor het grote publiek geldt vooral het privacy-aspect, met het naïeve verwijt dat zoekmachines persoonlijke ontboezemingen kunnen terugvinden, die men zelf aan publiek internet heeft toevertrouwd.

Anderszins is privacy wel degelijk serieus te nemen. Een paar weken geleden, bij een cursus, moest ik Google demonstreren. "www.google.com" en de beamer projecteerde meteen mijn persoonlijke zoekgeschiedenis van de afgelopen dagen. Bovenaan:

"aardappels afgieten" plassen

Plassen? Ja, in de betekenis van plasje doen. Mijn vader gebruikte daarvoor vaak de uitdrukking *de aardappels afgieten*. De vorige avond had ik me toevallig afgevraagd of dat een persoonlijk bedenksel was geweest, voortkomend uit beroepsmatige activiteiten - hij was aardappelhandelaar. Google gaf uitsluitsel. Nee: de uitdrukking kwam vaker voor. Waarvoor uiteraard "plassen" aan mijn zoekvraag moest toegevoegd om niet te verdrinken in aardappelrecepten.

Erg dat alle cursisten dit konden lezen, voor ik het snel wegklikte? Nee, zo privacygevoelig was het niet. En met weinig fantasie kun je veel pijnlijker situaties bedenken dan het brave woordje "plassen". Bovendien had ik het zelf in de hand gehad. Had ik maar niet - zonder beamer - tevoren mijn Gmail moeten checken. Identificatie bij Google had meteen een cookie op de PC gezet.

Maar het is wel degelijk bedreigend te bedenken dat na elke actie waarvoor identificatie nodig is, Google alle verdere handelingen onder jouw naam registreert en bewaart. Gestelde zoekvragen, maar ook welke gevonden pagina's je aanklikt en bekijkt. En zonder persoonlijke identificatie koppelt Google zulke gegevens aan IP-adressen van PC's.

Als je ook Gmail hebt, waarvan Google - voor jou persoonlijk - alle berichten full-text indexeert, en wanneer je nog een paar alerts hebt lopen, dan weet Google langzamerhand wel heel veel over je. En zoals ze webpagina's en zoekvragen kunnen analyseren voor inhoudelijke relevantiebepaling en om Adwords-reclame mee te sturen, zo kan uit al die persoonlijke gegevens een heleboel informatie over elke gebruiker worden afgeleid.

Google belooft alle persoonlijke informatie vertrouwelijk te houden en bazuint rond dat zij de rechtszaak hebben gewonnen, waarmee de Amerikaanse overheid hen wilde dwingen gebruikersgegevens beschikbaar te stellen. Maar wie garandeert dat dat altijd zo blijft, tegenover elk rechtssysteem en elke overheid? Blijf dus op uw hoede. Maar nu snel de aardappels afgieten, anders koken ze stuk.

* Google Blogoscoped Most Popular Words 2006 - blogoscoped.com/archive/2006-04-21-n53.html

De revelantie van dat alles

Search Engine Optimisers hebben intussen ook video op internet als werkterrein ontdekt. Bedrijven willen niet alleen dat hun website in zoekresultaten bij Google op één komt. Ook videootjes blijken een interessante bron voor reclame. In december beschreef een aflevering van SearchEngineWatch dan ook wat je moet doen om je video gevonden te laten worden en op te laten vallen. Toch had ik het gevoel dat dat nog erg op gewoon zoekmachinegebruik gebaseerd was, terwijl het bij video's meestal heel anders blijkt te werken. En niet alleen bij video's op YouTube of foto's op Flickr, maar eigenlijk bij vrijwel alle Web-2.0 diensten.

Je wordt geacht in tag clouds te kijken wat bij andere gebruikers de meest populaire tags zijn, in plaats van zelf te zoeken op basis van je eigen welomschreven informatiebehoefte. Je volgt andermans advies in plaats van zelf een zoekvraag te formuleren. Je hoopt met tags in Flickr wat leuke foto's te vinden, in plaats van gericht de klassieke plaatjeszoekers te gebruiken *(al is dat zoeken ook niet zo erg inhoudelijk, zoals Marten Hofstede in zijn laatste IP Weblog beschreef)*. Je volgt de suggesties van YouTube dat denkt te weten welke video's jij amusant zou kunnen vinden, in plaats van zelf te zoeken in de volledig doorzoekbare spraakherkende tekst van videoregistraties in Blinkx. Veel van die diensten zijn dan ook meer op verstrooiïng gericht dan op professioneel gebruik of studie.

Misschien moeten we maar het begrip **revelantie** invoeren om de kwaliteit van deze webdiensten te beoordelen. Als maat hoe goed ze in staat zijn om ons vet-cool amusement te onthullen, te reveleren. *Revelance ranking*, revelantie-ordening, zorgt voor een optimale volgorde, waarin de meest revelerende suggesties bovenaan komen.

Zoekend zou ik zelf nooit terecht gekomen zijn bij een zeven minuten durende video van een incident in een bibliotheek in Californië, waar een student met moslimuiterlijk met een elektrisch verdoofgeweer werd uitgeschakeld, omdat ie zich niet identificeerde. De video was opgenomen met een wild in het rond zwaaiend mobieltje, zodat je er haast niets op ziet, maar hij is wel al 561.873 keer bekeken (en na deze column nog wat vaker). Of een reclamevideo voor een bekend automerk, waarin domblondje bij een bibliotheekbalie op luide toon een patatje-mét bestelt. De dame achter de balie wijst er met uitpuilende ogen op dat dit een BIBLIOTHEEK is, waarna blondje verschrikt om zich heen kijkt en op meer gepaste fluistertoon haar bestelling herhaalt. Ga ik, als 136.536^{ste} kijker van die video, nu dat merk auto kopen?

Zo werkt dus revelantie. Meer amusementsbehoefte dan informatiebehoefte. Amusementsbehoeftenanalyse is kennelijk datgene waar SEO's zich op moeten richten. En Web-2.0 diensten moeten dan voor de revelantie zorgen.

De catalogus kan het wel alleen

"Zoekmachines kunnen het wel alleen" was de uitdagende titel die Theo Huibers had gegeven aan zijn bijdrage aan een studiedag over onderwerpsontsluiting begin april. Enigszins demagogisch suggereerde hij de verzamelde medewerkers van Wetenschappelijke Bibliotheken dat handmatig ontsluiten van informatie in toenemende mate overbodig wordt, dankzij moderne technieken van information retrieval. Wat Mike Lynch, directeur van Autonomy, enkele maanden geleden in een vraaggesprek met IP te berde bracht, had eigenlijk ongeveer dezelfde strekking. Goede zoekmachines (maar dan slimmere dan Google) waren volgens hem heel wat efficiënter in het produceren van goede zoekresultaten, dan systemen waarvoor hele ontologieën en thesauri moeten worden opgezet, de aanpak waarvan de bouwers van het Semantisch Web hun heil verwachten.

Een slecht verstaander zou hieruit kunnen concluderen dat we dus helemaal kunnen ophouden met het genereren van metadata voor onze catalogi. En dus eigenlijk onze catalogi kunnen opdoeken. Voor dat laatste zijn misschien best argumenten aan te voeren, maar dan niet vanwege de uitdagende uitspraken van Huibers en Lynch. Er is namelijk helemaal niet zo'n grote tegenstelling als die uitspraken lijken te suggereren. Voor redelijk gestructureerde catalogi blijven metadata en bepaalde vormen van autorisatie wel degelijk nodig. Maar computerprogramma's kunnen dergelijke metadata steeds beter zelf genereren. Daarvoor kunnen namelijk veel van dezelfde mooie technieken worden gebruikt, die achter de schermen zorgen dat de zoekmachines van Theo Huibers "het wel alleen kunnen".

Maar die mooie technieken, of ze voor de "geweldige" retrieval zelf zijn bedoeld, of voor het genereren van metadata, hebben veel meer digitale gegevens nodig dan nu in onze catalogi zitten. Ook Google wordt alleen maar steeds beter omdat het over zulke gigantische hoeveelheden data beschikt. Het is niet voor niets dat Google recent aankondigde in Amerika een telefonische "411"-inlichtingendienst te gaan aanbieden. Daarmee kan het grote hoeveelheden spraak oogsten, waarmee persoonsonafhankelijke spraakherkenning kan worden verbeterd, zodat ze ook een spraak gestuurde zoekmachine kunnen aanbieden. Grotere hoeveelheden digitaal beschikbare tekst - inhoudsopgaven van boeken, samenvattingen en flapteksten, zo mogelijk zelfs de hele inhoud - hebben we bovendien toch al nodig om catalogi te laten overleven. Want daarmee worden ook de extra diensten mogelijk, of we die nu Library-2.0 of Web-2.0 noemen, waarmee we nieuwe generaties gebruikers nog enigszins voor onze catalogi hopen te interesseren. Anders kunnen we die inderdaad wel opdoeken.

Na het verhaal van Huibers lijkt het opmerkelijk dat FRBR de laatste tijd - ook in IP - in de belangstelling staat. FRBR vormt namelijk de ultieme toepassing van zeer gestructureerde, onderling gerelateerde metadata beschrijvingen. Vooralsnog vraagt dat uiterst zorgvuldige menselijke inbreng. Maar wie weet, als daar wereldwijd maar genoeg in geïnvesteerd is, en ook genoeg digitale gegevens beschikbaar zijn, zullen zelfs zulke complexe catalogi "het" op den duur misschien wel alleen kunnen.

[27] juli 2007

Informatieinflatie: de verpakking én de inhoud

Zes jaar geleden heb ik het in een column (april 2001, column 12) ook al eens over informatie-inflatie gehad. Toen naar aanleiding van een Amerikaans onderzoek waarin was berekend dat onze informatieproductie ieder jaar verdubbelde (Lyman & Varian). Volgens mij zat het grootste deel van die verdubbeling alleen in de verpakking - terwijl YouTube toen nog niet eens bestond. Van een full-text platte tekst van 50 kB naar een video van 500 MB, waarin je diezelfde tekst ziet en hoort uitspreken, zoals ik onlangs persoonlijk mocht ondervinden. Daarmee kun je al dertien jaar verdubbeling realiseren, zonder dat je meer inhoud genereert. Bytes - of intussen liever exabytes - zijn dan ook geen goede eenheid om informatie in te meten.

Toch is het niet alleen de verpakking die voor de inflatie verantwoordelijk is. Op de studiedag over onderwerpsontsluiting, waarover ik twee maanden geleden al schreef (mei 2007, vorige column), hield bioinformaticus Barend Mons ons voor dat biomedische publicaties minder dan 10% nieuwe informatie bevatten; de rest is herhaling van wat al bekend is. Met weblogs lijkt dat minstens zo erg te zijn. In veel blogs bestaat meer dan 90% van de berichten uit herhalingen van dezelfde nieuwtjes. Met abonnementen op *feeds* van Search Engine Watch, -News, -Land, -Guide, -Journal, -Showdown, -Blog, -Herald, -War, -Roundtable en nog een heleboel andere, zie je elke scheet die Google laat, in veelvoud weerkaatsen en rondresoneren.

Search Engine Showdown en onze eigen IP Weblog horen tot de weinige uitzonderingen op de gewoonte om alle nieuwtjes van elkaar over te nemen. En bij nieuws over Web 2.0 en Library 2.0 gaat het al nauwelijks anders. De inhoud kent dus ook volop inflatie.

Hier lijkt behoefte aan een metaniveau, aan een superblog die alles samenvat. Maar dat lijkt juist het probleem te zijn. Bijna elke blog blijkt zichzelf die rol al te hebben toebedeeld, door alles te herhalen en de andere te citeren, zodat het eeuwig rond blijft zingen. Misschien dat virtuele communities, zoals bibliotheek20.ning.com, uiteindelijk wel in die behoefte kunnen voorzien. Daar is nog maar één plaats waar je alles vindt. Alleen ontbreekt de nodige structuur soms nog.

De komende vakantieperiode belooft enige adempauze te geven. Met dit ene nummer van IP zult u twee maanden toe moeten. En na terugkeer van vakantie, bevat uw persoonlijke Netvibes-pagina gelukkig alleen nog de tien of twintig meest recente posts van uw favoriete blogs. Als u uw PC weer aanzet, haalt uw lokale Feedreader alleen de posts van de laatste paar dagen op. Al gaat de inflatie van de inhoud bijna onverminderd door, u krijgt toch wat adempauze, doordat u even niet alles te zien krijgt. Die bloggers zelf lijken intussen nauwelijks aan vakanties en weekends te doen.

* Lyman & Varian / How Much Information? 2000 - http://groups.ischool.berkeley.edu/archive/how-much-info/

De Google-killer van 2008

Een vast thema van veel search-weblogs vorig jaar, was de vraag welke nieuwe zoekmachine zich zou ontpoppen als de ultieme Google-killer. Slimmere technologie en betere zoekresultaten, waar zelfs Google's PR niet meer tegenop zou kunnen. In veel van die berichtjes werd WIKIA als één van de kanshebbers genoemd. Niet dat daarvoor technische hoogstandjes werden aangekondigd. Jimmy Wales, de bedenker van de Wikipedia die hier achter zat, ging ons allemaal op wiki-achtige wijze inzetten, om te zorgen dat deze zoekmachine ons allemaal betere zoekresultaten zou leveren.

Eind december, nog net in 2007, was het zover. WIKIA kwam online. Maar in alle media zult u intussen gelezen of gehoord hebben dat alom grote teleurstelling heerst. Om technoweblog TechCrunch te citeren: *"Wikia Search is a complete letdown. ... [It] would be a disappointment even without the massive hype we've had to endure. And taking that hype into account, this product is an inexcusable waste of time."*

De conclusie dat WIKIA in 2008 zeker niet de grote Google-killer zal worden, lijkt dus geen al te gewaagde voorspelling. Maar wie wordt dat dan wel?
Ik weet een goede kanshebber: **Google zelf!**
Google die zichzelf om zeep helpt.

Google die probeert slimmer te zijn dan de gebruiker zelf. Die automatisch allerlei dingen doet, omdat ie denkt dat dat goed voor ons is.

Google die me op een zoekvraag naar wiki-software overspoelt met resultaten over de Wikipedia, omdat ie denkt dat ik dat wel bedoeld zal hebben. Wiki, wikipedia, allemaal één pot nat. NEE, als ik dat bedoeld had, dan had ik dat wel gevraagd.

Of de klacht van grootmeester Hans Ree in het schaaktijdschrift "Matten", dat hij op een zoekvraag naar de schaker Barendregt, wordt overspoeld met gegevens over de gemeente Barendrecht *(bron: Jos Damen)*. Anders gespeld, maar Google gaat er bij voorbaat al vanuit dat gebruikers niet kunnen spellen. En dat alles zonder mogelijkheid om aan te geven dat je niet van zulke ongewenste behulpzaamheid gediend bent. *[noot: dit bleek achteraf wel mogelijk]*

Als eerste stap op weg naar het uiteindelijke killen, zullen professionele gebruikers steeds vaker ontevreden worden, omdat zij zelf niet meer in de hand hebben waarnaar Google voor ze op zoek gaat. Zij zullen overstappen naar concurrerende zoekmachines die nog wel doen wat ze vragen. En hoe lang duurt het dan nog, totdat ook de gewone gebruiker te vaak merkt op het verkeerde been gezet te worden. Dat de loyaliteit van internet-zoekers uiterst vergankelijk kan zijn, heeft de geschiedenis van AltaVista geleerd. En het is haast onmogelijk je reputatie weer terug te krijgen, als intussen iedereen gebruik is gaan maken van .....
Ja van welke zoekmachine eigenlijk? De koning is dood, leve de koning, maar wie die nieuwe koning wordt, durf ik nog niet voorspellen. Maar wat ik wel weet: Google zal vast proberen die uiteindelijke killer zelf weer op te kopen.

De Yahoo!-killer van 2008

Zelf vind ik het altijd wat slapjes, columnisten die op hun vorige column voortborduren. Gebrek aan inspiratie, Sieverts? Maar dit keer is er aanleiding voor. Terwijl ik me vorige maand over de mogelijke dood van Google boog, leek vrijwel gelijktijdig de dood te worden aangekondigd van Google's enig overgebleven echte concurrent, Yahoo!. Zou overname door Microsoft echt het einde van Yahoo! betekenen?

Bij de allereerste IP-lezing, najaar 1999, was er een keynote van retrieval-goeroe Keith van Rijsbergen. In een interview vertelde hij dat Microsoft toen alle toonaangevende universitaire IR-onderzoekers naar zijn eigen IR-afdeling probeerde te halen. Desondanks zijn zoekmachines van Microsoft nooit echt iets geworden.

Om op de markt van Enterprise Search toch een rol te gaan spelen (en de concurrentie met specialist Autonomy aan te gaan), heeft Microsoft onlangs het Noorse FAST overgenomen. Navrant overigens dat FAST juist de motor was achter webzoekmachine AllTheWeb, die ooit als showcase voor FAST's zoeksoftware diende. En juist AllTheWeb werd enkele jaren geleden opgekocht door Yahoo!, dat nu .... inderdaad ook door Microsoft dreigt te worden opgekocht.

Op de webzoekmarkt heeft Microsoft's eigen Live Search ook al geen doorbraak bewerkstelligd. Of dat alleen een kwestie van gebrekkige marketing was - "Live" lijkt een wat vreemde merknaam voor een webzoekmachine - of dat de zoekresultaten ook slechter waren, laat ik even in het midden.

Wel blijft de indruk hangen dat de atmosfeer bij het bedrijf uit Redmond kennelijk verstikkend werkt op het innovatief vermogen van al die veelbelovende IR-specialisten die ze in dienst genomen hadden. Het blijft net of je naar MS-Office zit te kijken. Zelfs Yahoo! zou daarvan slachtoffer kunnen worden

Ben ik te eenzijdig? Een beetje wel. Er zijn ook heel speciale interfaces op Microsoft's Live Search, met fraai vormgegeven, uit echte filmfragmenten opgebouwde interactie. Miss Dewey bestaat al weer enkele jaren. Deze fraai ogende dame probeert je met gevatte (gesproken) opmerkingen en antwoorden tot zoekactiviteit te verleiden, maar is toch geen doorslaggevend succes geworden. LEFTvsRIGHT probeert, in het kader van de komende Amerikaanse verkiezingen, nu een politiek tintje aan het zoekproces te geven, maar lijkt ook nog nauwelijks populair.

Wordt Yahoo! werkelijk door Microsoft overgenomen? Het ziet er op dit moment even niet meer naar uit. Opmerkelijk ook dat gebruikers van Yahoo-diensten in het geweer kwamen, bijvoorbeeld die van Flickr. *"Fikken af van Flickr"*; daar kwam het op neer. Zulke persoonlijke identificatie met digitale diensten kennen we pas sinds internet en misschien pas echt sinds de opkomst van Web-2.0. Wie heeft zich ooit - anders dan om puur zakelijke redenen - druk gemaakt over overnames van commerciële diensten als Dialog, NewsEdge of ISI? Maar als Flickr in het geding is, borrelen gebruikers meteen over van inventiviteit en originaliteit, om de gehate overnemer belachelijk te maken. Maar killt dat Microsoft? Nou nee.


MS. DEWEY, JUST TELL ME
books


MS. DEWEY, JUST TELL ME
gun

Digitale kloven

Jan Tweepuntnul, onder zijn echte naam sinds kort ook redacteur van dit blad, publiceerde onlangs zijn 200ste blogbericht. Daarin sprak hij onder meer zijn twijfels uit of al die mooie nieuwe technieken waar we het steeds over hebben, zoals RSS, feedreaders, social bookmarking of Firefox add-ons, wel zo algemeen bekend zijn en zo veel gebruikt worden, zelfs door de jongere generatie. Hij is zeker niet de enige die vraagtekens zet bij de internet-vaardigheid van die zogenaamde generatie Einstein - de generatie waarvan ik intussen de grootvader had kunnen zijn.

Een rapport van de British Library en JISC meldde begin dit jaar ook al met nadruk dat het bestaan van wat daar de Google generation heet, een mythe is. Hoewel die generatie is opgegroeid met PC en internet (en zich niet kan voorstellen hoe het voelt om niet voortdurend online te zijn), blijkt ze gemiddeld helemaal niet handiger in het zoeken en vinden van informatie dan generaties die pas op latere leeftijd hebben kennisgemaakt met computer en internet.

Ook al hebben de opdrachtgevers voor dit onderzoek het gebruik van klassieke gecontroleerde informatiebronnen hoog in het vaandel staan, toch mogen we de auteurs zeker niet van vooringenomenheid betichten. Uit eigen ervaring kan ik het namelijk van harte eens zijn met zowel Jan als JISC. Als docent informatiemanagement merk ook ik dat de echte digitale kloof door alle generaties heenloopt.

Digitale vaardigheden betreffen niet alleen het gebruik van allerlei handige tools, maar ook het omgaan met de informatie zelf. Dat blijkt ook uit een artikel, onder de titel "In Google we trust", dat eind vorig jaar verscheen in het Journal of Computer-Mediated Communication. Het daar beschreven onderzoek liet zien dat universitaire studenten bijna blind afgingen op de positie van zoekresultaten in resultatenlijsten, zonder te kijken of de titels van gevonden webpagina's wel relevant klonken. Voor gestelde vragen niet-relevante titels op positie twee werden veel vaker aangeklikt dan zeer relevant klinkende op positie zeven.

Anderzijds meldde technoweblog TechCrunch onlangs juist dat in Amerika het "zoek-atheisme" in opkomst is - dat belooft nog wat voor de echte religie. Onder de kop "In Google we *no longer* trust" werden bevindingen van het Center for the Digital Future gerapporteerd. Nog maar de helft van de internetgebruikers blijkt resultaten en werking van zoekmachines zonder meer te vertrouwen, terwijl dat een jaar eerder nog 62% was.

Die laatste twee voorbeelden illustreren maar weer eens dat je bij elke opvatting wel een onderzoek kunt vinden, waarvan de uitkomsten die opvatting staven. En of je nu modern bent en RSS-feeds en Social Bookmarks gebruikt, of juist bij Netscape 3.1 bent blijven steken, voor het beoordelen van zulke uitkomsten blijft het toch vooral aankomen op eigen kennis, inzicht en kritisch vermogen (sprak opa).

* In Google We Trust: Users' Decisions on Rank, Position, and Relevance (Journal of Computer-Mediated Communication, 2007) - onlinelibrary.wiley.com/doi/full/10.1111/j.1083-6101.2007.00351.x

* Search Atheism On The Rise; In Google we no longer trust (TechCrunch 2008) - techcrunch.com/2008/01/17/search-atheism-on-the-rise/

Samen zoeken voor ons eigen

Tien jaar geleden heb ik in een column (mei 1998, column 6) in dit blad eens geschreven dat "Yahoo! won". Dat had toen niet betrekking op de zoekmachine, maar op de nu wat in vergetelheid geraakte onderwerpsdirectory. Die wijsheid over het winnen door Yahoo! had ik opgedaan op de Search Engine Meeting 1998. Ondanks de naam van dat congres, was het vertrouwen in zoekmachines behoorlijk aan het wegebben. Dat was nog net de pre-Google tijd, de periode waarvoor we maar een aparte jaartelling moeten invoeren, die aangeeft hoeveel jaar BG (Before Google) iets plaats vond. De bewering was toen dat zoekers liever zouden browsen in systematische indelingen dan met een zoekmachine te zoeken. Menselijke indexering, of eigenlijk classificering, zoals in Yahoo! of OpenDirectory, en ook in andere varianten, leek toen een betere oplossing voor onze zoekproblemen dan domme full-text zoekmachines.
Hoe anders kan het lopen!
Hoewel?

Ook nu is er weer ongenoegen over zoekmachines, zelfs - of misschien wel juist - over Google. Ook nu hoor je steeds meer klachten dat men te veel niet relevante informatie vindt. Of in elk geval niet-interessante, niet-betrouwbare, niet-kwalitatieve, niet-bedoelde informatie. Hoewel directories nu nauwelijks meer als oplossing van dit probleem worden gezien, komen er wel weer steeds meer systemen die diverse vormen van menselijke inbreng toepassen.

Of dat nu de Wikia zoekmachine van Jimmy Wales is, die op wiki-achtige wijze ons aller oordeel over webpagina's wil laten meewegen in de te presenteren resultaten. Of dat nu een zoekmachine als Stumpbleupon is, of misschien Mahalo als exponent van de "human factor" waarvoor Danny Sullivan al de benaming Search-4.0 heeft bedacht. Of wellicht Microsoft's meest recente wanhopige poging op zoekgebied, "SearchTogether", waarmee je via een browser plug-in je eigen zoekvraag samen met anderen kunt proberen op te lossen.

Ook dit is niet allemaal even nieuw. Met Web-2.0 is recommendatie, het ontdekken wat anderen je aanbevelen, via tag clouds en most popular items, vaak ook al belangrijker dan zelf zorgvuldig zoeken.

Zo zijn we ooit van zoeken naar browsen en van browsen weer naar zoeken gegaan. En zo gaan we in het spoor van Web-2.0 van zoeken naar aanbevelen. Deze trend suggereert dat de computer het nog altijd niet helemaal zelf kan. Dat er mensen nodig blijven om een echt oordeel te geven. Maar misschien is ook dat maar tijdelijk. Het semantisch web doet beloften dat we de computer ook kunnen leren te bepalen waar iets over gaat en wat iets betekent. En hem ook laten beoordelen of iets betrouwbaar is en kwalitatief de moeite waard. Alleen is het nog even de vraag wanneer dat semantisch web een keertje klaar zal zijn. Misschien dat ik daar in een column over nog eens tien jaar op terug kan kijken.

Interoperabel erfgoed

"Interoperability" begint een beetje een modekreet te worden. Wel eentje waarover zelfs Engelse native speakers hun tong breken als ze het lekker snel willen uitspreken. Dat bleek op een congres eind juni bij de Koninklijke Bibliotheek, dat onder die titel was aangekondigd, maar in werkelijkheid geheel gewijd was aan het Europeana-project. Onder het motto "Connecting Cultural Heritage" moet daarin een portal gerealiseerd worden voor geïntegreerde toegang tot Europese digitaal erfgoed collecties.

Om al die collecties in één keer te kunnen doorzoeken en ook verdere uitwisseling mogelijk te maken, is interoperabiliteit van al die verschillende systemen natuurlijk een essentiële voorwaarde. Toch viel op die bijeenkomst niet veel inhoudelijks te leren over de nieuwste ontwikkelingen op dat terrein. Het ging vooral om de presentatie van een werkend prototype van het nieuwe zoeksysteem. Daarbij was ook tijd ingeruimd voor discussies, waarin wijzigingen aan het systeem konden worden voorgesteld en besproken, zelfs door deelnemers die verder niet bij het project betrokken waren.

Wat op het congres wel bleek, was dat interoperabiliteit in de praktijk nog een moeizaam proces is. In de serie "De Standaard" die eerder in dit blad is verschenen, kwamen verschillende veelbelovende technieken aan de orde, die bij interoperabiliteit een belangrijke rol zouden kunnen spelen: RDF (oktober 2006), SKOS (november 2006) en OWL (december 2006). Al die mooie technieken ten spijt, blijkt men in de praktijk toch nog te moeten terugvallen op het handmatig opzetten van concordanties, zelfs al op het niveau van voor beschrijvingen gebruikte velden.

Nu is dat bij dit project ook niet zo gek, omdat voor het eerst alle uithoeken van het cultureel erfgoed hierbij worden betrokken: musea, archieven, bibliotheken en audiovisuele collecties. Dat tussen zo uiteenlopende soorten organisaties nog nauwelijks gemeenschappelijke standaarden worden aangehouden voor het (digitaal) beschrijven van het zeer uiteenlopende materiaal, is natuurlijk niet zo verwonderlijk.

Wat wel opvalt is dat geautomatiseerde semantische technieken voorlopig nog tekort schieten om al tijdbesparing te kunnen realiseren bij het koppelen en integreren van zulk heterogeen materiaal. Ook toepassing van een algemene ontologie als Cidoc-CRM blijkt hiervoor geen panacee. Deze vanuit de museumwereld ontwikkelde ontologie, beoogt meer een methodologisch dan een technisch hulpmiddel te zijn. Het heet dan ook een "conceptual reference model". Hoewel algemeen van opzet, vindt het voorlopig nog alleen toepassing in gespecialiseerde omgevingen binnen beperkte onderwerpsdomeinen.

Ondanks mijn kritische woorden over nog tekort schietende technieken, is dit onderdeel van de European Digital Library natuurlijk wel degelijk een interessant en uiterst ambitieus project. Helaas mag ik u het URL van het prototype nog niet doorgeven, omdat het systeem nog niet robuust genoeg is om de zoekacties van al die geïnteresseerde IP-lezers al lekker snel te kunnen verwerken. Meer informatie en een voorgeprogrammeerde demo-tour op www.europeana.eu.

Broeikas of ijstijd

Kennisklimaatverandering is het motto van deze OCN (Online Conferentie Nederland). Dat lijkt nog niet meteen gepaard te gaan met het verschijnen van een deltarapport van een daartoe ingestelde deltacommissie. Zelfs laat deze kreet in het midden of het hier om opwarming zal gaan of juist om een nieuwe ijstijd. Of bibliotheken broeierige broedplaatsen voor nieuwe activiteiten worden of dat ons een koude winter te wachten staat, met verder teruglopende gebruikscijfers en krimpende budgetten. Ook zegt dit motto weinig concreets over de inhoud van de lezingen die de bezoekers van de OCN te wachten staat.

De summiere gegevens op de OCN-website maakten het ook al niet makkelijk om vooraf duidelijk inzicht te krijgen wat die inhoud zal zijn. Toch was uit de beschrijving van de verschillende "tracks" uiteindelijk wel een globaal beeld te krijgen: het thema van ruim driekwart van het programma lijkt direct of indirect samen te hangen met leren en onderwijs. Hoe integreren bibliotheken en informatiecentra hun activiteiten met die van onderwijs en onderzoek? Hoe leveren ze daarvoor het materiaal aan? Wat is hun rol in digitale leeromgevingen? En hoe ondersteunen ze - ook buiten het onderwijs zelf - de lerende organisatie? Onontkoombaar valt daarbij natuurlijk ook de term *competentiegericht* onderwijs.

Welke infrastructuur is voor dat alles nodig? En hoe wordt verder de informatiegeletterdheid van zowel scholier als werknemer verbeterd?

Dat woord *informatiegeletterdheid* - als letterlijke vertaling van "information literacy" - bekt misschien niet zo lekker als *mediawijsheid*, maar ik geef daar eigenlijk de voorkeur aan. Het geeft beter aan waar het om gaat, om een vaardigheid en niet om een intellectuele verworvenheid als *wijsheid*. Een wijsgeer hoef je (gelukkig) allerminst te worden om met nieuwe media en informatiebronnen te kunnen omgaan.

Ook onderwerpen als gaming vallen onder dat thema "leren". Op de masterclass "Library-2.0", enkele weken geleden in Rotterdam, gaf Stephen Abram al aan dat computergames uitstekende middelen zijn om kinderen te laten leren. Zelfs als het niet zo heel inhoudelijk is, traint het nog wel het geheugen - ook bij volwassenen trouwens. En dat geheugen blijven we nodig hebben, want het is een illusie te denken dat je, in een tijd dat je alles te allen tijde kunt opzoeken, niets meer zou hoeven weten en onthouden. Een voldoende voorraad parate kennis blijft een essentiële voorwaarde om al die opgezochte of via RSS binnenkomende informatie in de juiste context te plaatsen en te kunnen beoordelen. Alleen moet die parate kennis wel ergens zijn opgedaan.

Laat bibliotheken in het kader van die aangekondigde kennisklimaatverandering maar broeikassen worden, waar levenslang leren en kennisdelen tot snelle wasdom kunnen komen. Dat broeikaseffect vereist gelukkig geen extra CO-2 uitstoot. Laat uw SUV dus thuis en kom gerust met trein of fiets naar de OCN.

Crisis

"Stagnerend IBL belangrijkste aanjager van kredietcrisis!"
Hoe dat zo? Het bleek dat banken elkaar niet meer vertrouwden, waardoor het InterBancair Leenverkeer vrijwel tot stilstand was gekomen. Ook de NVB kon daar in een gemondialiseerde markt weinig aan doen. De NVB? Ja, de Nederlandse Vereniging van Banken.

Dat andere IBL - het "onze" - gaat gelukkig een stuk beter. In een lezing op de OCN extrapoleerde Mark Deckers de huidige Zoek&Boek IBL-aanvragen van de OB Deventer naar heel Nederland. Zo kwam hij uit op een potentieel jaartotaal van 2 miljoen aanvragen. Het is natuurlijk de vraag of dat met ongewijzigd beleid nog altijd te betalen blijft bij gemiddelde verwerkingskosten van €30 per aanvraag - het kostte de zaal wat moeite Mark dit bedrag te ontfutselen.

Dat getal 30 werd in de direct daaropvolgende OCN-lezing van Bas Savenije nog een keer genoemd, maar nu als een bedrag in dollars. Wetenschappelijke uitgevers vragen gemiddeld $30 voor een PDF van een enkel artikel, als jouw organisatie of jij zelf geen digitaal abonnement op dat tijdschrift hebben. De dollarkoers staat laag, maar dit is nog altijd niet echt goedkoop. Voor die IBL-aanvraag wordt tenminste nog wat gedaan. Dat PDFje zou zonder extra kosten aan veel meer mensen kunnen worden aangeboden als het maar niet zo duur was.

Maar anderzijds is het misschien ook wel goedkoop. Want door licenties en wetgeving worden UB's nu nog gedwongen via IBL aangevraagde digitale artikelen eerst uit te printen en dan weer te scannen, voordat ze - nu als digitaal image - naar de klant gemaild mogen worden. Dan is het misschien nog wel goedkoper om die $30 voor de klant te betalen.

Overigens floreert dat boeken-IBL waar Mark Deckers het vooral over had, via bibliotheekcatalogi. Dat is wel opmerkelijk, want onze huidige catalogi zijn eigenlijk volstrekt ongeschikte hulpmiddelen om erachter te komen in welke boeken - non-fictie - een bepaald onderwerp aan de orde komt. Vrijwel alleen als een boek in zijn geheel gewijd is aan een gevraagd onderwerp (dat dan meestal ook wat breder is), zijn de kale metadata uit onze catalogi enigszins afdoende om het daarop te kunnen vinden. Welke specifieke onderwerpen in afzonderlijke hoofdstukken en paragrafen van een boek behandeld worden blijft onbekend, zodat die kennis in de praktijk onvindbaar blijft, hoe mooie zoekinterface je ook bouwt.

Daarvoor moeten we dan toch weer op Google wachten. Op dezelfde OCN lichtte de Gentse bibliothecaris Sylvia Van Peteghem een klein tipje van de sluier op, hoe haar cont(r)act met Google Books tot stand gekomen was. Wat je daar ook over mag denken, een feit is dat met dit soort technieken eindelijk de echte inhoud van boeken gevonden kan worden. IBL is dan bijna niet meer nodig om je kenniscrisis op te lossen.

Meiregen maakt dat ik slimmer word

Zoals meiregen maakt dat ik groter word, zo maakt Google dat ik dommer word. Treffende overeenkomst tussen deze twee uitspraken is dat ze niet erg "evidence based" zijn. Toch was die tweede uitspraak de kern van een essay, deze zomer in The Atlantic Monthly: *"Is Google Making us Stupid; what the internet is doing to our brains".* Daarin beschrijft Nicholas Carr dat hij geneigd is alleen nog oppervlakkig feitjes te verzamelen en zich geen tijd meer te gunnen diepgravender teksten te lezen.

Ter verklaring beroept Carr zich op Marshall *(the-medium-is-the-message)* McLuhan, die in de jaren '60 al postuleerde dat de media geen passieve informatiekanalen zijn. Zo denkt Carr dat de wijze waarop informatie op internet wordt verspreid, ook grote invloed heeft op de manier waarop zijn eigen hersens met informatie omgaan: *"Once I was a scuba diver in the sea of words. Now I zip along the surface like a guy on a Jet Ski".* In zijn betoog staat Google uiteraard metaforisch voor alle zoekactiviteiten waarmee we gegevens op internet bijeen sprokkelen.

Het woordje "stupid" uit de titel van dit essay lijkt op nogal gespannen voet te staan met uitkomsten van recent hersenonderzoek, die ook al in IP's nieuwsrubriek zijn gemeld. Aan de hand van fraaie MRI-plaatjes werd bij UCLA aangetoond dat mensen van middelbare leeftijd juist veel meer van hun brein gebruiken bij het zoeken op internet, dan wanneer ze gewoon lezend informatie tot zich nemen. Ook voor mij is er dus nog hoop.

De discrepantie tussen deze twee opvattingen doet denken aan recente discussies over de multitasking capaciteiten van de huidige jeugd. In de NRC hield Marjoleine de Vos een vooral gevoelsmatig betoog, dat het maar schijn is dat de internetgeneratie dat zo goed kan, tegelijk muziekluisteren, MSNchatten, huiswerkmaken en videoclipkijken. Op Frankwatching zette Joost Steins Bisschop dit betoog weg als absolute onzin. Wow!ter en anderen verwezen op hun beurt weer naar het boek "Het puberende brein" van Eveline Crone en naar onderzoek van de Amerikaanse psychiater Hallowell, die aantonen dat multitasking in feite "sequential tasking" is, waarbij jongeren door versnipperde aandacht per taak beduidend minder leren dan bij langduriger aandacht. Dit verschilt eigenlijk niet erg van hetgeen de bijna middelbare Carr bij zichzelf waarnam.

Ook hier zorgt de metaforische Google dat ik in korte tijd diametraal tegengestelde gegevens kan verzamelen, die allemaal even goed onderbouwd lijken. En intussen houdt Google goed in de gaten waarnaar ik zoek en welke resultaten ik bekijk om daar rekening mee te houden bij het presenteren van resultaten van nieuwe zoekacties, zodat ik nog sneller nog meer oppervlakkige feitjes kan verzamelen. Ook al maakt Google ons misschien dommer, door ons zoekgedrag helpen wij Google in elk geval slimmer te worden.

* Nicholas Carr / Is Google Making Us Stupid? What the Internet is doing to our brains (2008) - https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/

* UCLA study finds that searching the Internet increases brain function (2008) - http://newsroom.ucla.edu/releases/ucla-study-finds-that-searching-64348

Worlds apart

Weinig lezers van dit blad zullen onkundig zijn van het feit dat XML voor ons vak van belang is. Bij de Nederlandse XML-club, voluit nog de "SGML/XML User Group Holland", kom je toch maar weinig bezoekers uit onze kring tegen. Nu moet ik bekennen dat ik zelf ook pas vorige maand voor het eerst op hun jaarcongres kwam, en dan nog vooral vanwege de uitnodigende congrestitel "Semantiek en Interoperabiliteit".

Dat bezoek bevestigde het beeld van aparte werelden, die van de *hard core* XMLers en die van de bibliothecarissen. Met als gevolg dat in onafhankelijke evolutie meer keren hetzelfde wiel wordt uitgevonden. Daarin verschillen we niet van de natuur. In de evolutie - 2009 is Darwin-jaar - is wel dertig keer opnieuw "het oog uitgevonden". Lees "The ancestor's tale" van Richard Dawkins daar maar op na. Elke keer een beetje anders van constructie en werking, maar wel met steeds dezelfde functie.

Zo gaven medewerkers van de Londense Royal Academy of Music op het XML-congres een lezing over het model dat ze hadden bedacht voor het beschrijven van de levensloop van muziekwerken, met plaats voor arrangementen, variaties, uitvoeringen, ervaringen (bij uitvoeringen), registraties, uitvoerende orkesten en dirigenten, concertzalen, alles aan alles gelinkt. Als ik dat zo opschrijf voel je al aan dat dat erg kan gaan lijken op het FRBR-model waar we onze nieuwe bibliotheekcatalogi op gaan baseren. Toch was het idee onafhankelijk ontstaan; pas achteraf waren ze achter het bestaan van FRBR gekomen.

Bovendien werd dit model niet gebaseerd op simpele relaties tussen entiteiten zoals bij FRBR, maar gebruikte men Topic Maps. Aan die op XML gebaseerde methode van ontsluiting wordt in "onze" wereld juist weer erg weinig aandacht besteed. De topics waar het daar om draait, zouden wij misschien concepten noemen, maar dan wel op een heel speciale manier gekarakteriseerd en toegepast. Bij een topic kunnen *names* horen, allerlei benamingen - ook in meer talen - om dat topic te omschrijven. Ook kan een topic met een *type* worden beschreven, zodat je de concepten in soorten kunt onderverdelen.

Onverwacht is dat een topic bovendien *occurrences* kan hebben, wat inhoudt dat het ook gekarakteriseerd wordt door de informatie-items die erover gaan. Bij documenten horen hier dus geen trefwoorden, maar andersom torsen trefwoorden documenten met zich mee. Verder zijn er ook nog *associations* waarmee relaties tussen topics kunnen worden aangegeven. Het meest bijzondere aan dit alles is dat die *names*, *types*, *occurrences* en *associations* van een topic zelf ook weer topics zijn en zelf dus ook weer met *names*, *types*, *occurrences* en *associations* kunnen worden beschreven, die op hun beurt ook weer .....

Inderdaad, een slang die in zijn eigen staart blijft bijten. Dat moesten die aparte werelden misschien ook wat vaker met elkaar doen.

Informatieanalfabetisme

Laatst raakte ik op een feestje in gesprek met iemand die jarenlang bij een technisch bedrijf had gewerkt dat met een grote multinational in het zuiden des lands gelieerd scheen te zijn. Toen ik probeerde uit te leggen wat mijn werk inhield, ontmoette ik groot onbegrip. Hij had geen idee wat een informatiespecialist was, waarom je er een nodig zou hebben en wat het nut van externe informatie zou kunnen zijn. Als je iets wilde weten, ging je gewoon bij de concurrent langs en probeerde je erachter te komen hoe zij een probleem oplosten of een bedrijfsproces hadden ingericht, door slim rond te kijken, door strikvragen te stellen of door stiekem een stopwatch te gebruiken. Persoonlijk zou ik zulke technieken eerder bedrijfsspionage dan informatiemanagement noemen.

Is dergelijk informatieanalfabetisme beperkt tot obscure technische bedrijfjes in het zuiden des lands? Bij de overheid schijnt het al niet veel beter te zijn, als je het gerucht mag geloven dat voor de wetgeving rond het paddoverbod gewoon een lijst mogelijk hallucinogene zweefzwammen uit de Wikipedia was gekopieerd. Al heb ik dat niet persoonlijk gecontroleerd, ik vond wel bevestiging van iemand die de lijsten metterdaad had vergeleken. Deze twee voorbeelden van aninformatisme illustreren dat er inderdaad werk aan de winkel is voor de "Taskforce Arbeidsmarktcommunicatie Informatiespecialisten" van Peter Evers en Kees Westerkamp. Hun poging het vak van informatiespecialist beter op de kaart te zetten, verdient zeker alle steun.

Maar die specialisten moeten dan ook gecertificeerde kwaliteitsinformatie kritisch blijven bekijken. Volgens het vorige nummer van IP zou ik veel te weinig salaris ontvangen. Het gerapporteerde onderzoek van de Amerikaanse Special Libraries Association (SLA) wees uit dat de gemiddelde Europese informatiespecialist € 73.000 verdiende - tegen maar ruim € 50.000 voor de gemiddelde Amerikaan. Zelf dacht ik altijd redelijk aan de bovenkant van mijn beroepsgroep te zitten, maar nu bleek mijn inkomen naar Europese maatstaven juist ver te zijn achtergebleven. Voor al die andere lezers van IP die dus kennelijk zo veel meer verdienen, kan het abonnementsgeld dan wel verhoogd worden, zodat ik eindelijk voor mijn columns betaald kan krijgen.

Of zou er toch iets mis zijn? Uit meer gedetailleerde gegevens op de SLA-website bleek dat de vermelde bedragen waren gebaseerd op gegevens van bijna 3000 Amerikanen en maar liefst 26(!) Europeanen. Waren die 26 veelverdieners echt representatief voor alle vakgenoten uit zo'n 40 Europese landen? Verder leken er nog wat rekenfoutjes te zijn gemaakt bij het omrekenen van de verschillende valuta's tussen de detailgegevens, het SLA persbericht en het IP-berichtje. Maar ook de zeer exacte € 64.986 uit de oorspronkelijke gegevens liggen nog altijd ver boven de bedragen uit Erwin la Roi's beloningsmonitor. Zo blijken informatieprofessionals, of dat nu bij SLA is of bij IP, ook wel eens informatieamateurs te zijn.

Groter en kleiner

Voor een te geven cursus heb ik me weer eens meer dan gebruikelijk in webzoekmachines verdiept. Moet "machines" echt nog in het meervoud? In Nederland krijg je sterk de indruk dat alleen nog Google bestaat. Diens marktaandeel is nog net geen 100%, maar veel scheelt dat niet meer. Voor de gezamenlijke concurrentie resten nog maar luttele procenten. Dat verschilt hemelsbreed met Google's thuisland waar alleen al Yahoo! een marktaandeel van 20% heeft behouden. Dezelfde Yahoo! die in Nederlandse taartpuntdiagrammen niet eens meer zichtbaar is - want ver onder de 1% gezakt.

Toch is dat heel gek. Uit oppervlakkig onderzoek merkte ik namelijk dat Yahoo! voor de meeste zoekvragen intussen meer - en soms zelfs veel meer - oplevert dan Google. Niet dat kwantiteit het enig belangrijke argument is in een situatie van informatieovervloed. Toch willen we, als we voor onze organisatie toegang tot een commerciële database aanschaffen, daar graag de dekking van weten. Dat is nog altijd een belangrijke factor bij de afweging tussen twee concurrerende producten. Merkwaardig dat we dat op het web ineens niet meer van belang achten. In Nederland vertrouwen we Google op zijn blauwe ogen. Een verhouding van marktaandelen tussen Yahoo! en Google van minder dan 1:100, terwijl Yahoo! voor de meeste vragen een betere dekking blijkt te bieden. Het is oneerlijk gesteld in de webzoekwereld.

Hierbij moet je wel de vraag stellen hoe betrouwbaar de vaak gigantische aantallen zijn die zoekmachines binnen een seconde voor elke zoekvraag op ons scherm toveren. Die vraag is makkelijk te beantwoorden: volstrekt niet! Daar wordt maar een slag naar geslagen, op basis van de eerst binnenkomende gegevens uit een paar kleine stukjes van hun indexen. Ze hoeven voorlopig toch maar tien resultaten te laten zien. Alleen voor die gevallen waar het aantal resultaten onder de duizend uitkomt, valt uiteindelijk te verifiëren hoeveel echt gevonden is. Dat aantal bleek gemiddeld maar de helft van wat aanvankelijk gemeld werd, met af en toe een uitschieter naar zelfs een factor tien minder. En ook in die controleerbare gevallen bleef Yahoo! gemiddeld nog duidelijk winnen.

Opmerkelijk was ook hoe sterk de relatieve resultaten van zoekvragen tussen de verschillende zoekmachines in drie jaar tijd waren veranderd. Vond Exalead op een bepaalde Nederlands/Friestalige zoekvraag drie jaar geleden nog half zo veel als Google, nu was dat nog maar 2%, terwijl voor andere vragen de relatieve opbrengst juist verdubbeld was. Vond Gigablast op "poldermodel" drie jaar geleden maar 3% van wat Google vond, nu was dat 30%. Voor diezelfde vraag was het resultaat van MSN-Live zelfs van 8% tot 67% van Google's aantallen gestegen. Kortom webzoeken blijft een dynamisch en onvoorspelbaar gebeuren. En cursisten moet weer goed worden ingepeperd dat ze alles steeds weer moeten uitproberen.

Ik tagde, jij tagde, ...

Ik tag, jij tagt, hij tagt, wij zouden graag getagd hebben.
Het heeft misschien even geduurd, maar tagging lijkt helemaal geaccepteerd door de professionals in ons vakgebied. Het is helemaal Bibliotheek-2.0. Alleen zijn we er alweer te laat mee, naar het schijnt.

Pas recent kwam ik er toevallig achter dat de toonaangevende Web Technology Blog ReadWriteWeb vorige zomer de *tag cloud* al dood verklaard heeft. Inclusief een cartoon met grafsteen:

The Tag Cloud, born 2002, died 2008.

In een latere blogpost, eind december, ging RWW daar nog een keer op in. De problemen van ongecontroleerde tagging en folksonomies, waar de bibliotheekwereld steeds tegenaan gehikt had, werden vanuit de hoek van de webtechnologen nu ook ineens breed uitgemeten. Collaborative tagging en folksonomies zag men niet meer als *de toekomst van het web*.

Dat verhaal werd opgehangen aan een bespreking van twee hulpmiddelen die moeten helpen de betekenis van tags wel goed vast te leggen: *semantic tagging services*. Faviki doet dat door aan webpagina's die je wilt taggen, die woorden en begrippen te ontlenen die ook voorkomen in een uit de Wikipedia gedestilleerde lijst begrippen. Uit dat rijtje voorstellen kun je kiezen, zodat je wel gedwongen wordt vastgelegde termen te gebruiken.

Zigtag lijkt dat te doen door pagina's die door gebruikers van dezelfde tag zijn voorzien, te clusteren in groepjes waarbinnen die tag vermoedelijk in dezelfde context is bedoeld. Hoe zinnig dat gebeurt, werd mij uit een paar testjes niet meteen duidelijk. In elk geval lijkt op die manier niets te worden gedaan aan het probleem van het gebruik van synoniemen.

De onbetrouwbaarheid van wat verschillende mensen met hun tags bedoelen - en waaraan Zigtag en Faviki beweren iets te doen - werd door RWW overigens niet meer als het enige nadeel van tagging gezien. Nee, RWW verwacht dat mensen inherent te lui zijn om steeds te blijven taggen bij alles wat ze op het web doen, bij elke blogpost, bij elk bookmark, bij elke foto en elke video. Na een paar jaar begint de lol er al weer af te gaan.

RWW verwacht dat tagging alleen haalbaar blijft, als er systemen komen die niet zo zeer onze tags begrijpen, maar die ze helemaal automatisch voor ons toekennen. Waar hebben we dat eerder gehoord? Automatische verrijking, classificatie en thesaurering: de heilige graal van de informatieprofessional. Die professional, misschien niet te lui, maar wel als duur gezien, blijft het belang van metadateren wel inzien. Met goede metadata leg je betekenis wel meteen vast. Metadateren en taggen zijn dus heel verschillende zaken. Tenzij je met dat woord "taggen" natuurlijk juist "metadateren" bedoelt. Ik metadateer, jij metadateert, wij metadateren met metadata die we tags noemen (zolang dat woord nog hip is).

Allemaal aan de dope

In Wageningen werd 1 april een symposium over Bibliometrie gehouden. Aanleiding daarvoor waren de citatieanalyses die waren uitgevoerd voor de beoordeling van diverse Wageningse onderzoeksgroepen. Met analyses van de impact van gepubliceerde artikelen is de Wageningse Universiteitsbibliotheek lokaal de concurrentie aangegaan met het Leidse Centrum voor Wetenschaps- en Technologiestudies (CWTS) van Ton van Raan. Over de daarbij toegepaste methodieken heeft Wouter Gerritsma in oktober 2006 al eens in dit blad geschreven.

Net zoals Search Engine Optimizers allerlei technieken propageren om websites hoger te laten scoren in Google, zo kregen eveneens bij het symposium aanwezige wetenschappers meteen adviezen hoe ze individueel, als hele onderzoeksgroep of zelfs als hele universiteit hoger konden scoren in de internationale citatiewedloop. Elkaar veel citeren was er daar één van. In dat verband was het aardig dat net in de maart-aflevering van ResearchTrends, een tweemaandelijkse nieuwsbrief over bibliometrische analyse en trends in wetenschappelijk onderzoek, aandacht werd besteed aan het sociale aspect van citatiegedrag. Bestaan er citatieclubs, waarbinnen vriendjes uit sociale netwerken elkaar vaker citeren dan anderen? Daar schenen geen aanwijzingen voor te zijn. Zij het dat daartoe aangehaald onderzoek al weer twee tot zes jaar oud was.

Behalve aan SEO adviezen, deden de aanbevelingen voor citatie-optimalisatie me ook denken aan adviezen om als bibliotheek je catalogus-records in Google te krijgen - wat diezelfde Wageningse bibliotheek trouwens ook lukt. Je klanten vinden je dan tenminste nog op Google. Maar wat als iedereen dat doet? Dan vind je een gezocht (of ongezocht) boek 683x in Google, uit de catalogi van alle 683 verschillende bibliotheken die dat boek bezitten en allemaal hun catalogusrecords in Google hebben weten te krijgen. Dan heeft niemand daar meer iets aan, ook de bibliotheek niet die zich zo bij zijn gebruikers wilde profileren. Zo heeft ook niemand er meer voordeel van als iedereen alle trucs van stal haalt om zijn citatiescore met technische middelen te verhogen, en niet door beter onderzoek te doen en betere artikelen te schrijven. Het is als met wielrennen: als iedereen dope slikt en EPO spuit, haalt niemand daar meer winst uit. Maar tegelijk kan niemand zich permitteren ermee op te houden.

Gelukkig leveren sommige van de genoemde methoden van citatiedoping ook objectief voordeel op, doordat ze zorgen dat artikelen een grotere kans krijgen te worden gelezen. Ruimere verspreiding van kennis en wetenschap is immers een edeler doel dan het bevredigen van universitaire boekhouders die cijfertjes willen om op basis daarvan geld te verdelen. Of gaat het lezen van die gepushte artikelen weer ten koste van andere artikelen? Zaten potentiële lezers al aan het maximale quotum van wat ze 7x24 kunnen lezen? Misschien kun je ook wel wat slikken om per dag meer kilokarakters te kunnen verwerken.

RT @sieverts http://is.gd/zpv0


In een verhaal over bestrijding van informatie-overload, deed ik onlangs nog de uitspraak dat Twitter dat probleem alleen maar erger maakte.
about 1 ago from ip-twitter


Toch bleek ik zelf niet zo principieel; nog geen 2 weken later nam ik mijn al meer dan een jaar slapende twitteraccount ook maar in gebruik.
about 1 ago from ip-twitter


In mijn lezing had ik Andrew Goodman wel aangehaald, die in februari (http://is.gd/ldS1) meldde minder te gaan bloggen en meer te twitteren.
about 1 ago from ip-twitter


Zijn argument was dat de tweets van de groep mensen die hij op twitter volgde, hem hielpen om "sneller te denken". Ja, dat wilde ik ook wel.
about 1 ago from ip-twitter


Inderdaad is er al verandering merkbaar; op het NLbiliobloggersfront is het rustiger; fervente bloggers houden zich soms een week lang stil.
about 1 ago from ip-twitter


@Jantweepuntnul blogde daar ook al over: door het vele twitteren houdt hij geen stof (of tijd?) meer over om ook nog blogposts te schrijven.
about 1 ago from ip-twitter


Maar dat is meteen de merkwaardige interne contradictie: die blogpost van Jan is een lang doorwrocht verhaal, vol mening, over- en afweging.
about 1 ago from ip-twitter


Het was zelfs langer dan de 450 woorden die deze columns mogen tellen, zoiets lukt je nooit binnen de 140 tekens plus spaties van een tweet.
about 1 ago from ip-twitter


Anderzijds is Twitter natuurlijk prima voor korte berichten, meldingen, nieuwtjes en attenderingen, liefst met bijbehorende ingekorte URL's.
about 1 ago from ip-twitter


Maar of ik dan ook geïnteresseerd moet wezen in de treinvertragingen van @ritanila en @gbierens, of in muziekjes die @zbdigitaal beluistert?
about 1 ago from ip-twitter


Zoals ik ook zwemdiploma's van @Wowter's kinderen op de koop toe moet nemen, om interessantere meldingen van al die mensen te kunnen volgen.
about 1 ago from ip-twitter


In mijn lezing had ik dus wel gelijk: de verdunning van informatie neemt zo steeds verder toe, en ook Clay Shirky krijgt steeds meer gelijk.
about 1 ago from ip-twitter


Het probleem waar we mee zitten is geen informatie overload, het is filter-failure, ons onvermogen de krenten nog uit de info-pap te vissen.
about 1 ago from ip-twitter


Daardoor helpt Twitter bij mij nog niet echt bij dat beloofde snellere denken, of zouden mijn eigen beperkte vermogens daarvan oorzaak zijn?
about 1 ago from ip-twitter


Een ander effect van twitteren is dat het nogal eens moeite kost om alle over te brengen informatie in de beschikbare 140 tekens te stoppen.
about 1 ago from ip-twitter


Bij de groeiende groep bibliotwitteraars zie ik nog geen SMS-taal, maar soms wel geheimtaal alleen begrijpelijk voor een groepje ingewijden.
about 1 ago from ip-twitter


Dat komt ook doordat tweets nogal eens worden gebruikt als een tussenvorm tussen publieke mededeling en persoonlijk gerichte e-mail of chat.
about 1 ago from ip-twitter


Daardoor lijdt Twitter als algemene informatiebron, veel sterker dan andere bronnen die ik volg, aan een zekere mate van ontoegankelijkheid.
about 1 ago from ip-twitter


Wie dat wellicht ook voor deze aflevering van de column vindt gelden, kan heel goed gelijk hebben; hij bestaat namelijk helemaal uit tweets.
about 1 ago from ip-twitter


Bovendien heb ik ze zo geformuleerd dat elk van deze 20 tweets op exact 140 tekens uitkwam, zodat het meteen echte twitteratuur is geworden.
about 1 ago from ip-twitter

Dit product is GEEN zoekmachine

Het lijkt een nieuwe trend te zijn hoe nieuwe zoekdiensten zich proberen te profileren. Je noemt jezelf GEEN zoekmachine. Twee recente voorbeelden zijn uitgebreid in het nieuws geweest: Wolfram|Alpha en Bing.

De claim van Wolfram|Alpha het NIET te zijn, was terecht. Het is geen zoekmachine, maar een dienst die poogt gebruikersvragen te analyseren - wat overigens niet altijd lukt. En als W|A een vraag begrepen denkt te hebben, worden concrete antwoorden afgeleid uit allerlei tamelijk gestructureerde datacollecties. Dus niet uit willekeurig gevonden webpagina's. Helaas heeft het niet erg geholpen dat ze zeiden een "computational knowledge engine" te zijn, want op allerlei plaatsen - zelfs in IP - werd W|A toch geafficheerd als Google-concurrent op Google's eigen webzoekterrein.

Ook Microsoft's Bing wordt heel bewust GEEN zoekmachine genoemd. Een peperdure marketingcampagne benadrukt dat Bing een "decision engine" is, wat ze daarmee ook mogen bedoelen. Deze claim geen zoekmachine te zijn, is voor een belangrijk deel onzin en grootspraak. Dat vrijwel niemand die claim serieus neemt, verheugt mij dus. Want Bing onderscheidt zich niet wezenlijk van andere zoekmachines. Net als bij Google en Yahoo! krijg je nog gewoon webpagina's voorgeschoteld.

Toch is er wel wat aan de hand. Maar dan niet alleen met Bing, maar met alle grote zoekmachines. Ze zijn steeds minder recht-toe-recht-aan zoekmachines die reproduceerbaar resultaten opleveren, die exact aan je zoekvraag voldoen.

Steeds meer proberen zoekmachines indirect de mogelijke bedoeling af te leiden die de gebruiker met zijn vraag heeft. Zo willen ze ook op slecht gespecificeerde vragen betere antwoorden kunnen geven. Een mooi en lovenswaardig streven. En nog mooier als ze daar ook in slagen.

Toch heb ik wel een bedenking. Het maakt zoekacties veel minder reproduceerbaar en controleerbaar. En dat moet professionele zoekers wel degelijk zorgen baren. Informatieprofessionals leren in hun opleiding dat informatieonderzoek, ook al noem je het "deskresearch", een vorm van onderzoek is. En dat vereist dat het doelgericht, controleerbaar en dus ook reproduceerbaar wordt uitgevoerd. En dat in gevoelige gevallen het zoekproces achteraf ook volledig verantwoord kan worden. Dat nu, wordt op het web steeds lastiger.

Je weet steeds minder goed wat zoekmachines met je zoekwoorden doen. Of dat hetzelfde is als gisteren en of dat morgen ook nog het geval zal zijn. Wat op een .nl versie anders gebeurt dan op een .com versie. Of resultaten zijn weggelaten - of naar positie 1000 gedegradeerd - omdat de zoekmachine op basis van eerder gedrag vermoedde dat je die niet bedoeld zult hebben. Voor gewone gebruikers prima als dat op slechte vragen minder slechte antwoorden oplevert. Maar voor professionele zoekers niet, zeker als het op goede vragen minder goede antwoorden geeft. Jammer genoeg zijn wij de enigen die merken dat zoekmachines zo wat minder zoekmachine worden.

Data doen er weer toe

Vaak wordt de indruk gewekt dat data eigenlijk niets voorstellen, dat het gaat om de informatie die we daaruit kunnen destilleren. En zelfs dat bleek de afgelopen jaren niet goed genoeg. Uiteindelijk zou het alleen nog maar draaien om de kennis die we op basis van die informatie kunnen opbouwen. In een soort raffinageproces krijg je steeds hoogwaardiger producten: van data naar informatie naar kennis.

Toch begon het heilige geloof in kennismanagement de laatste tijd weer wat af te brokkelen. En zelfs lijken we nu terug te keren naar de "data". Eigenlijk ook niet zo gek. Studies, publicaties, teksten - informatie kortom - zijn in de meeste gevallen, in oorsprong gebaseerd op gegevens. Als we dat, enigszins wetenschappelijk, primaire onderzoeksdata noemen, dan was daarvoor al enige tijd aandacht bij DANS (Data Archiving and Networked Services), een dienst van de KNAW, waar wetenschappers hun kostbare data voor de eeuwigheid bewaard en toegankelijk kunnen maken. Alleen ligt die eeuwigheid voor veel betrokkenen nog wel erg ver weg.

Intussen wordt ook bij universiteitsbibliotheken het belang ingezien van primaire onderzoeksdata. Als producten van de wetenschap moeten niet alleen publicaties toegankelijk zijn, maar liefst ook de primaire gegevens waarop die gebaseerd zijn. In Utrecht is daarom een proefproject gestart waarin onderzoeksgroepen hun primaire gegevens kunnen opslaan en zelf kunnen bepalen wie ze daar toegang toe geven. Daarin is Utrecht intussen zeker niet meer de enige.

Software voor dit project was ook gewoon bij Harvard University beschikbaar, waar ze al een paar jaar langer ervaring hiermee hebben.

Ook SURF is recent een werkgroep begonnen die hier onderzoek aan gaat doen. Is het Utrechtse project in eerste instantie vooral gestart vanuit de behoefte van onderzoeksgroepen om in hun virtuele kenniscentra ook onderzoeksgegevens beschikbaar te hebben, voor SURF is het een logische uitbreiding van de repositories waarin universiteiten en hogescholen publicaties van hun medewerkers en studenten opslaan, maar waarin nu ook "complexe documenten", inclusief achterliggende datasets beschikbaar moeten komen.

Misschien nog opmerkelijker dan deze voorbeelden, is dat ook in het kader van het semantisch web "data" de kop opsteken. Het "semantische" aan dat nieuwe web gaat namelijk juist om betekenis en dus om kennis. Toch heeft Tim Berners Lee zelf het begrip *linked data* ingevoerd. In een interview op de ReadWriteWeb blog noemt hij dat een *grass root* beweging. Niet van bovenaf door W3C opgelegd, maar door allerlei eigenaren van veel informatie aan de basis, wordt hier gewerkt aan het koppelen en navigeerbaar maken van grote hoeveelheden gegevens. Toch is juist daarvoor nog wel iets meer nodig. Data kun je alleen zinvol koppelen als je er de betekenis van weet, liefst vastgelegd in metadata. Kennis van de betekenis is dus essentieel. Zo komen data en kennis dus toch nog bij elkaar.

Het einde van ....

Het lijkt mode de dood of het einde van instituties en verschijnselen te verkondigen. Zo leek het einde van de bibliotheek aanstaande als je de ophef mocht geloven rond een *Prep School* in de buurt van Boston die al zijn 20.000 boeken had opgedoekt. De quote waarmee de *headmaster* die stap verantwoordde was dan ook wel om in te lijsten: "When I look at books, I see outdated technology". Toch blijkt op veel plekken het tegendeel. Zo werd een dezer dagen in Utrecht de nieuwe "Universiteitsbibliotheek Binnenstad" geopend. Daar zijn klassieke gebouwen omgetoverd tot een ruime en lichte omgeving waar vele kilometers van die *outdated technology* in open opstelling klaar staan. En als zelfs de CNN website uitgebreid aandacht besteedt aan "The future of libraries, with or without books" (ja, dat dan weer wel) lijkt er niet veel mis.

The death of search. Met die kop trok marketing weblog Frankwatching begin september de aandacht. Maar daar zijn web-marketeers en SEO specialisten aan het woord, die vrezen hun waren niet meer zo makkelijk via zoekmachines aan de man te kunnen brengen. Een eerste tegenwerping is al dat het volume aan zoekvragen op internet tussen juli 2008 en juli 2009 met ruim 40% blijkt te zijn toegenomen (en bij Google zelfs met bijna 60%). Maar hoe zit het met dat *real time web* dat klassieke zoekmachines obsoleet zou maken?

Voor het merendeel van professionele informatiezoekers - voorzover ze niet alleen met marketing bezig zijn - is het maar marginaal om alleen te zoeken in hetgeen de afgelopen paar minuten op Twitter of Facebook is gebabbeld, om te zien wat nu hypes en trends zijn. Google CEO Eric Schmidt maakte zoeken intussen weer hot met zijn recente uitspraak dat zoekmachines in de toekomst rechtstreeks aan onze hersens gekoppeld zullen zijn. Maar dat vind ik toch wat luguber klinken.

De dood van de weblog schijnt ook aanstaande. Niemand heeft meer tijd weloverdachte blogposts te schrijven, omdat continu maximaal 140-letterige berichtjes de wereld ingestuurd moeten worden. Miljoenen enthousiast begonnen weblogs blijken inderdaad niet meer te worden bijgehouden. Misschien maar goed ook. Dat helpt tegen de homeopathische informatieverdunning waarover ik al eerder klaagde. Mits alleen de beste blogs overblijven. Lijden aan twitterblock schijnt trouwens al weer de nieuwste aandoening te zijn.

En dan heb ik het nog niet eens gehad over de dood van Yahoo! dat zijn zoekziel aan Microsoft verkocht heeft, maar toch beweert met zijn zoekinterface met Bing te willen concurreren. En ook niet over de dood van RSS, waarvan ik overigens alleen maar las dat het NIET dood zou zijn. Wat een onzin al die doodsberichten: we leven als IP-ers toch juist in een ongelooflijk levendige - en dus interessante – tijd.

[45] november 2009

Let's get instant satisfaction

Recent attendeerde Lorcan Dempsey in zijn weblog en via Twitter nadrukkelijk op het dit voorjaar al uitgekomen rapport "Discoverability" van de Universiteit van Minnesota. Dat signaleert vijf belangrijke trends waarmee bibliotheken rekening moeten houden. Voor lezers die met hun digitale tijd meegaan, waren daar geen echte eye-openers bij. Toch wil ik "trend #2" hier even noemen: *"Users expect discovery and delivery to coincide"*.

Dat komt aardig overeen met wat ik eerder dit jaar in een online bijdrage op de IP-website schreef. In "De mythe van de catalogus" had ik het over de *instant satisfaction* die webzoekmachines - lees Google - aan zoekers bieden. Dat heeft onze klanten verwend. Er is een verwachtingspatroon ontstaan van onmiddellijke bevrediging van informatiebehoeften. De ontdekking van informatie en de aflevering ervan in je browser vallen bij zoekmachines vrijwel samen. Daar kunnen onze klassieke catalogi nooit tegenop, wat voor mooie trucjes men ook bedenkt om mensen ietsje sneller aan boeken of papieren copietjes te helpen.

Met de komst van het Open Archive Initiative zijn deelnemende universiteiten, hogescholen en onderzoeksinstituten wel al aardig richting *instant satisfaction* geschoven. Via een overkoepelende zoekmachine als OAIster kun je de meeste publicaties uit hun meer dan 1100 lokale repositories met enkele klikken full-text op je scherm krijgen. Des te opmerkelijker dat OAIster recent is overgenomen door OCLC, de kampioen van de uitgestelde bevrediging via bibliotheekcatalogi.

Als ik het goed heb begrepen, is het zelfs de bedoeling om OAIster met zijn 23 miljoen full-text gelinkte publicaties te integreren in Worldcat, de gewone catalogus met zijn 90 miljoen titels. Een vreemd mengsel, ook al is OCLC de werkgever van die Lorcan Dempsey die het belang van *instant delivery* zo benadrukt.

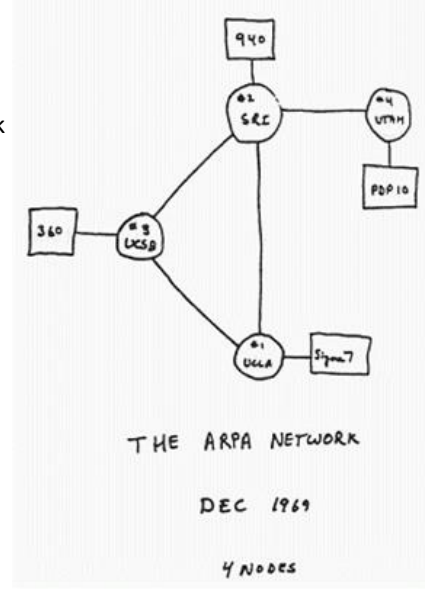
Dat de overname van OAIster op discussielijst NGC4LIB - over de *"next generation catalog"* - voor enige reuring zorgde, is niet verwonderlijk. Daar was men vooral bang gegevens uit OAIster niet meer in andere diensten te mogen blijven gebruiken. Ook internationaal staat OCLC namelijk niet echt bekend om het vrijgeven van metadata aan al die bibliotheken die, door hun catalogiseeractiviteiten, zelf gezamenlijk die metadata geproduceerd hebben. Zo'n "gijzeling" van de eigen catalogusgegevens is ons in Nederland ook niet geheel onbekend.

Door de open aard van het "harvesting"-protocol voor de OAI metadata, lijkt dat gevaar bij OAIster afwezig. Zo is er al enige tijd een Zwitsers alternatief, "Scientific Commons" dat dezelfde OAI-metadata oogst en doorzoekbaar maakt als OAIster doet. En wie maar wil, kan zelf vergelijkbare of juist heel andere diensten op diezelfde metadata gaan bieden. Het zou mooi zijn als dat ook eens met de metadata uit onze catalogi zou kunnen. Ook al bieden die nog geen onmiddellijke bevrediging. Of het E-book dat ooit wel zal kunnen bieden is weer een heel andere discussie.

* Discoverability: Investigating the Academic Library's Changing Role in Connecting People to Resources - https://conservancy.umn.edu/handle/11299/116586

* De mythe van de catalogus - sieverts.pbworks.com/f/mythe-catalogus.htm

Jubileren



Normaal houd ik er niet van om met Oud & Nieuw goede voornemens te formuleren of jubelend achterom te kijken. Nu de jaren '10 van deze eeuw al weer zijn ingegaan, laat ik deze houding toch maar even varen in mijn eerste column van dat nieuwe decennium. Bovendien was 2009 een jaar van allerlei jubilea.

Zo was er aandacht voor het 40-jarig bestaan van internet. Het plaatje met de vier knooppunten van Arpanet, waarmee het december 1969 allemaal begon, dook allerwegen op.

Hoewel het woord "hypertext" nog ouder is dan internet - al in 1963 door Ted Nelson geïntroduceerd - vierden de hypertext markup language en het hypertext transfer protocol - de basisingrediënten van het web - pas hun 20ste verjaardag. Voor de VU reden om Tim Berners-Lee dit najaar een eredoctoraat te verlenen.

Wat dat web aanvankelijk voor Nederland betekende, wordt duidelijk als je de Dutch Home Page van 15 jaar geleden nog eens bekijkt. Alle 40 (!) Nederlandse websites pasten nog overzichtelijk op een kaartje van Nederland. Lang geleden in een column (nr 19, mei 2003) vertelde ik al eens dat dat plaatje alleen nog maar bestaat, omdat van die website een afbeelding in de NRC had gestaan. Een plaatje dat ik met vooruitziende blik had uitgeknipt. Kijk maar hierboven als je die Dutch Home Page anno 1994 nog eens op krantenpapier wilt zien, want digitaal is hij nergens meer. Een jaar later was die kaart al een stuk voller (als afbeelding gevonden in een congresbijdrage over de DHP uit die tijd).

Begin 1995, nu dus 15 jaar geleden, schreef ik ook mijn eerste artikel over webzoekmachines. Dat was in LOGIN, één van de bladen waaruit IP is voortgekomen. Lycos was toen de eerste serieuze zoekmachine om full-text in websites te kunnen zoeken.

Met zijn 1,5 miljoen (!) webpagina's had Lycos toen een aanzienlijk deel van het web geïndexeerd. Hoe webzoekmachines en hun crawlers te werk gingen was nog helemaal nieuw voor ons informatieprofessionals. Als je dat artikel nog eens naleest, merk je hoe ingrijpend de zoektechnieken zich sindsdien hebben ontwikkeld. Niet zo gek als je van 1,5 miljoen webpagina's naar vele honderden miljarden moet meegroeien. Ook niet zo gek dat van Lycos alleen de naam nog over is. Achter de huidige zoekmachine met die naam zit nu gewoon Yahoo!. En dat is nu net de enige andere naam die nog over is van de andere, meer rudimentaire zoekmachines uit begin 1995.

U verwacht nu ook een voorspelling voor het komende decennium? Ik begin er niet aan! Want toen ik 15 jaar geleden over Lycos schreef, had ik niet kunnen voorzien wat we nu allemaal met internet, zoekmachines, sociale media of twitter kunnen doen. Wel wil ik een ieder interessante jaren '10 toewensen.

* Zoeken op WWW met Lycos (LOGIN 1995) - sieverts.pbworks.com/f/lycos.pdf

Een integratiedebat

Wat ik deed op het moment waarop ik hoorde dat Kennedy vermoord was, weet ik echt niet meer. Maar ik weet nog wel heel precies waar ik was op *nine-eleven*. Bij de UB Amsterdam, met een gezelschap bibliothecarissen uit het hele land. Daar presenteerden enkele leveranciers systemen, voor wat we nu "federated search" noemen, maar toen nog gewoon "metasearch" heette. In de theepauze vertelde een collega me ontzet dat een vliegtuig het WTC was binnengevlogen. Aanvankelijk weigerde ik haar te geloven; het leek te onwaarschijnlijk om waar te kunnen zijn. Hoewel we acht jaar geleden het nieuws nog niet direct konden checken met I-phone of ander mobiel gerief, werd toch snel duidelijk dat mijn ongeloof ten onrechte was.

Dat ik die middag ook al geen geloof hechtte aan de claims van de presenterende systeemleveranciers, bleek achteraf een stuk legitiemer. Dat hun metasearchsystemen de ultieme oplossingen waren voor content-integratie - met één uniform zoekinterface in een heleboel verschillende content tegelijk kunnen zoeken - daarvan wisten ze mij niet echt te overtuigen. De meeste andere aanwezigen waren aanzienlijk enthousiaster. Maar die hadden ook nog geen eigen zoeksysteem zoals wij in Utrecht. Wij hadden net een eerste versie van "Omega", waarmee in mijn ogen op veel elegantere manier geïntegreerd in content van al even diverse leveranciers werd gezocht.

Acht jaar later wordt mijn scepsis veel algemener gedeeld, zowel door aanvankelijk enthousiaste kopers van die metazoeksystemen, als door sommige leveranciers van toen. Op basis van de Open Source zoekmachine Lucene, worden steeds meer systemen ontwikkeld voor "integrated search", het met moderne zoekmachinesoftware doorzoekbaar maken van metadata die afkomstig zijn van allerlei tijdschriftuitgevers en informatieleveranciers. Op dezelfde manier als dat in Utrecht in Omega gebeurt.

Zo hebben de bibliotheken van Delft en Tilburg het Meresco-systeem geïmplementeerd, dat artikelen uit wetenschappelijke tijdschriften van Elsevier (en binnenkort van meer uitgevers) integreert met de eigen catalogus. Zo heeft het Israëlische ExLibris - maker van één van die "oplossingen" van acht jaar geleden - nu een soortgelijk systeem te koop onder de naam Primo. En dat kun je zelfs bij ExLibris laten hosten. Een soortgelijke uitbestede dienst wordt geleverd door het Amerikaanse Serials Solutions, onder de naam Summon. Een oplossing waarmee de Rotterdamse Universiteitsbibliotheek een pilot is gestart. Wat de beste vorm van dit soort integratie is, leidt nauwelijks meer tot debat.

Blijft het hierbij? Ook landelijke samenwerking en landelijke oplossingen voor "integrated search" lijken op komst, als onderdeel van een nieuwe Nationale Informatie Infrastructuur. Dat is zelfs in het beleidsplan van de KB doorgedrongen, zoals blijkt uit het interview met Bas Savenije, elders in dit blad. Zal ik me later als bejaarde ook kunnen herinneren wat ik deed op de dag dat ik hoorde dat *alles* in de nieuwe NII geïntegreerd was?

Alle buzz waait over

Twee columns lang had ik de naam Google al niet meer laten vallen. Dat had ik nog wel een tijdje zo willen houden, maar in ons vak blijkt dat te moeilijk. Je krijgt steeds weer te maken met iets waar ook Google bij betrokken is. En dat betreft lang niet altijd Google's core business - het zoeken. Overigens wordt terecht betwijfeld of voor *big G* zoeken nog echt de kern is. In een interview in Vrij Nederland stelde Siva Vaidhyanathan dat Google eigenlijk primair een advertentiebedrijf is. Maar wel eentje met slimmere technieken om advertenties aan de man te brengen dan enige andere speler op die markt.

Toch heb ik het idee dat Google een beetje van zijn voetstuk aan het vallen is. En dan niet alleen vanwege de kritische houding van Vaidhyanathan in VN. En ook niet alleen vanwege mijn eerdere klachten dat je niet meer weet wat je aan Google hebt als je ermee aan het zoeken bent (juli 2009, column 42). Want die klachten heb ik bij Bing nog veel sterker. Nee, het is het hele beeld.

Je weet nu ook niet meer wat je aan Google hebt als je Gmail gebruikt. Ongevraagd en onaangekondigd probeerde Buzz ineens onze Gmail binnen te dringen. Wie zat op Buzz te wachten? Wie had meteen door wat het voor je privacy betekende, als je niet wilde achterblijven en meteen Buzz activeerde? Wie horen we trouwens nog over zinvolle toepassingen van Google Wave? Die vorige Google-hype, waarvan ik al in geen weken mijn berichten meer heb bekeken, omdat al snel een onoverzichtelijke weinig chronologische chaos vol onduidelijke zijpaadjes ontstond.

In de redactieblog van InformatieProfessional zette ik ook vraagtekens bij het nut van de vindbaarheid van verder onzichtbaar blijvende boekinhoud in Google Books. Die ontoegankelijkheid van veel boekinhoud is zo strijdig met de *instant satisfaction* van onze informatiebehoefte, waaraan we juist gewend geraakt zijn door webzoekmachines als Google. Oké, ik weet dat Google het wel zou willen laten zien, als er geen auteursrechtelijke wetten in de weg stonden en praktische bezwaren (laat staan weemoedigheid die niemand kan verklaren).

Google Books brengt me weer terug bij Siva Vaidhyanathan, door Vrij Nederland als 'mediaprofessor' gekarakteriseerd. In december bij de 'Society of the Query' in Amsterdam had hij met zijn indrukwekkende gestalte ook al indruk gemaakt met zijn kritische woorden over Google. In VN wijst hij onder meer op het probleem dat er geen enkele garantie voor continuïteit is, als je het aan een oncontroleerbare en ontransparante speler als Google overlaat de bibliotheek van de wereld te worden. Zelfs bij universiteitsbibliotheken was het enige tijd mode te zeggen dat je niet moest proberen met Google te concurreren. Akkoord, financieel en technologisch kan dat inderdaad niet, maar volledig afhankelijk zijn van *big G* is nog heel iets anders.

Opstanding

Dit nummer van uw vakblad valt rond Pasen in de bus. Associaties met "opstanding" leiden mij wat oneerbiedig naar de veelvuldig doodverklaarde OpenDirectory. De meeste onderwerpsgidsen hebben het trouwens moeilijk. Wie gebruikt nog de Yahoo-directory? De OpenDirectory heeft ook nog tegen dat zijn naam en URL verschillend zijn. Het is het Open Directory Project (ODP) op de site DMOZ.org. Daarin staat DMOZ voor "Directory Mozilla", dezelfde Mozilla die achter Firefox en Thunderbird zit. Toch is de directory al jaren in verval: weinig nieuwe aanwas (al zes jaar een omvang van 4,5 miljoen sites), tekort aan vrijwilligers (naar Wikipedia overgelopen?) en weinig bekendheid bij gebruikers.

Maar er zijn tekenen van wederopstanding. Weblog SearchEngineLand publiceerde oktober vorig jaar een uitgebreid verhelderend interview over DMOZ. Een week later haakten ze daarop in met een voorstel hoe DMOZ aan nieuwe editors kon komen, weer ouderwets "web librarians" genoemd. En vorige maand verschenen berichten dat eind maart een vernieuwde versie van de directory live zou gaan. Als u dit leest, zou die er dus al moeten zijn.

Nog meer tekenen. Google is één van de vele zoekportals waarop de OpenDirectory als onderwerpsgids wordt aangeboden. Maar in Google's dienstenoverzichten kwam hij al een paar jaar niet meer voor, zodat je een zoekmachine nodig had om uit te vinden waar Google hem op zijn eigen site verborgen had. Totdat enkele weken geleden de Directory ineens weer prominent opdook als één van de dertien diensten onder "even more". Althans in de Engelse versie, want zoals gebruikelijk loopt de Nederlandse Google daar maanden op achter.

En op het Duitse eBay werd in februari het opnemen van een site in DMOZ zelfs bij opbod verkocht. Of duidde het lage eindbod van 87 euro juist nog op gebrek aan belangstelling?

Naast tekenen van wederopstanding, zijn er namelijk ook ontwikkelingen die blijvend belang van onderwerpsgidsen dubieuzer maken. Plotseling duikt het begrip "site-less web" op. Begin februari zag ik die kreet voor het eerst in een weblog van Paul Gillin. Op het huidige sociale web is geen eigen website meer nodig om je publiek te bereiken. Met diensten als Twitter, Facebook, Flickr, Delicious en Posterous kunnen personen én bedrijven hun doelgroepen soms makkelijker bereiken dan met statische eigen websites. Zelfs Associated Press kondigt zijn nieuwe berichten op Twitter aan en verwijst daarin niet naar het eigen AP.org domein, maar naar hun Facebook-pagina's. Volgens sommigen een visionair idee. Hoezo site-less? Facebook is toch ook een site! Ja, maar niet de eigen site.

Mijn vraag is nu hoe onderwerpsgidsen als DMOZ (en dus ook de Yahoo-directory en onze Nederlandse Startpagina) daarmee zullen omgaan. Hun "ontsluiting" van het web is primair site-gebaseerd. Welke ingangen moeten zij rubriceren als het site-less web verder doorzet? Ondanks de paastijd lijkt DMOZ' definitieve wederopstanding nog niet zo'n uitgemaakte zaak.

Primaire reacties

Een collega die in de digitale wereld als Wowter bekend staat, attendeerde me op een artikel dat binnenkort in het Journal of Academic Librarianship gaat verschijnen. De *corrected proof* bleek al digitaal beschikbaar. Reden mij hierop te attenderen waren de titel "Perspectives on Primary and Secondary Sources" en het feit dat ik altijd de begrippen primaire, secundaire en tertiaire informatiebronnen introduceer bij mijn inleiding op de VOGIN-cursus. Althans mijn persoonlijke interpretatie daarvan, begrijp ik na lezing van het artikel. Dat concludeert namelijk dat mensen daar vaak heel verschillende dingen onder verstaan. Nu was dat al niet meer nieuw voor me. Collega docenten van de voormalige archiefschool bleken ook een andere definitie van primaire bronnen te hanteren dan ik. En in een project voor het opslaan van "primaire onderzoeksdata" waarbij ik in Utrecht betrokken ben, staat dat "primair" ook veel dichter bij die archiefbetekenis dan bij mijn eigen oorspronkelijke opvatting.

In het genoemde artikel had een Amerikaanse bibliothecaris aan onderzoekers, docenten en studenten gevraagd wat zij onder primaire en secundaire bronnen verstonden. In zijn bibliotheekinstructie wilde hij namelijk aansluiten bij vigerende opvattingen. Hoewel bleek dat die per discipline en doelgroep uiteenliepen, kwam hij toch tot een eigen beslissing. Primaire informatie is voor hem onderzoeksdata en archiefstukken, terwijl artikelen waarin die gegevens (primair) worden beschreven, voor hem al secundaire bronnen zijn. En auteurs die over die secundaire publicaties van andere auteurs schrijven, genereren zelfs al tertiaire bronnen.

Help! Wat moet ik dan met mijn bibliografische databases die ik altijd secundaire informatiebronnen noem, de bronnen waarin gegevens over oorspronkelijke artikelen - *mijn* primaire bronnen - worden samengebracht? En mijn tertiaire bronnen, de verzamelplaatsen van mijn secundaire bronnen, de bibliografieën van bibliografieën, databasedirectories en overzichten van onderwerpsgidsen?

Tot overmaat van ramp werd vervolgens op een SURF onderzoeksdata-forum een model voor onderzoeksdata gepresenteerd, dat met niveaus werkt. Daarbij komt level 0 aardig overeen met de primaire bronnen van mijn archiefcollega's. Maar daar verschenen ook volgende niveaus, waarbij de primaire onderzoeksdata die beta-onderzoekers als stromen digitale getallen uit hun meetapparaten krijgen, eigenlijk al op level 1 of 2 terecht zouden moeten komen. Zijn die in de ogen van historici dus eigenlijk al secundair of tertiair? En dan de artikelen daarover? En dan de artikelen over die artikelen? En dan de databases waarin die artikelen zitten? En dan ... Het lijken de Drosteblikjes wel.

Als informatiezoeker houd ik bij mijn karakterisering van digitale informatiebronnen nog maar even vast aan mijn eigen definities van primair en secundair. Als ik *mijn* primaire bronnen eigenlijk al secundair of misschien wel tertiair zou moeten noemen, dan worden mijn tertiaire bronnen intussen al quartaire of misschien wel quintaire of sextaire bronnen. Mijn cursisten zijn dan vast al afgehaakt. Maar zoals u begrijpt is dit een nogal primaire reactie van mijn kant.

* Primary and Secondary Sources (Journal of Academic Librarianship, 2010) - https://www.sciencedirect.com/science/article/abs/pii/S0099133310000650

Content voor uw P-book reader

Op de TechCrunch technoblog stond recent een bericht over PediaPress (keurig met -i-a- gespeld) dat in samenwerking met de Wikimedia Foundation de mogelijkheid biedt om van een eigen selectie van Wikipedia-pagina's een boek te laten maken. Een boek dat echt gedrukt wordt, op papier, een zogenaamd P-book.

Dacht ik met dat "P-book" een originele grap bedacht te hebben. Op internet werd ik snel uit de droom geholpen: op www.urbandictionary.com bleek al vier jaar een LOL-definitie van P-books te staan:
"Semi-portable macroscopic devices into which information and stories were typically inscribed by means of inks printed into "paper". Primarily made from trees and various plants and fibers, and commonly found glossed and polished on the outside or wrapped in cowhides. P-books were considerably weighty depending on the amount of information stored inside them, and despite their size usually contained only one title!!"

Maar terug naar het onderwerp: zo'n initiatief verbaast me. Zijn we niet tevreden met al die *digitally born informatie*? Waar je overal, waar je maar wilt, online toegang toe kunt krijgen. Die in principe full-text doorzoekbaar is. Doen we met zijn allen niet stinkend ons best (Google en KB voorop) om alle interessante oude papieren boeken eindelijk juist digitaal beschikbaar te krijgen? Dan is dit toch de omgekeerde wereld. Toepassen van dode-bomen-technologie.

Nu moet ik toegeven dat ik zelf boter op mijn hoofd heb. Zoals sommige lezers misschien al hadden opgevangen, zijn we met een gezellig groepje van vier collega's bezig geweest een nieuw boek te schrijven. Dat gaat over wat ik in ouderwets bibliotheek-jargon maar kortweg *ontsluitingstechnieken* noem. Voor het samenstellen van dat boek hebben wij ook een wiki gebruikt. Nog wel een beetje hybride. Sommige hoofdstukken hebben we in de wiki zelf geschreven, maar andere gewoon als Word-document dat na elke revisie opnieuw op de wiki werd gezet. En ja, ondanks die digitale wiki-voorgeschiedenis, wordt ook dit toch weer een echt papieren boek - zij het dat uitgever Biblion, misschien nog niet helemaal van harte, ook wel een digitale versie uitbrengt.

Nu dat nieuwe boek van ons toch eenmaal ter sprake is gekomen, laat ik hier dan ook de titel maar meteen onthullen. Die luidt 'Organiseer je informatie: aan de slag met thesauri, taxonomieën, tags en topics'. Dat klinkt hopelijk wat uitnodigender dan het toch wat saaie 'Woordsystemen', de titel van het boek dat een soort voorganger hiervan was. Een "soort" voorganger, want in dat nieuwe boek komen ook heel andere onderwerpen aan de orde, zoals het structureren van websites en ook nog iets over ontologieën en over die mooi allitererende T-begrippen uit de ondertitel. Vorige week was de definitieve tekst klaar; in september moet het verkrijgbaar zijn. Inderdaad als P-book en als E-book.

GII: Een G van Gemeenschappelijk of van Global

Op 1 juni was een gemengd gezelschap in Utrecht bijeen om over ontwikkelingen rond de GII geïnformeerd te worden. De GII? De Acronymfinder kent daar wel 68 betekenissen van, van *Global Instability Index* tot *Ghana Integrity Initiative*. En dichterbij de in Utrecht bedoelde I's ook *Government*, *Geospatial* en zelfs *Global Information Infrastructures*. Maar waar het 1 juni echt over ging was de Gemeenschappelijke Informatie-infrastructuur.

In zijn inleiding schetste Bas Savenije het vergezicht van een gemeenschappelijk systeem, waarin fysieke en digitale collecties gezamenlijk voor iedereen doorzoekbaar worden. De fysieke collecties uit de klassieke GGC - het Gemeenschappelijk Geautomatiseerd Catalogiseersysteem - doorgespoeld naar een wat moderner Open Index en aangevuld met de catalogusgegevens uit de OB-sector. Dat alles via WorldCat ook automatisch doorgesluisd naar Google. En die digitale collecties dan? Gegevens van artikelen, ook van wetenschappelijke uitgevers, komen in een voorlopig nog aparte index, onder de werktitel Panorama. Via één of meer portals - Bibliotheek.nl maakt er nu al eentje - moet dat alles integraal en vooral ook eenvoudig doorzoekbaar worden.

De wetenschappelijke artikelen die voor veel mensen nu nog achter hoge tolmuren zitten, worden daarbij ook voor iedereen toegankelijk. Voor sommige gebruikersgroepen gratis op basis van centraal betaalde licenties, voor degenen die daar niet onder vallen tegen redelijker tarieven dan nu. Gebruikersidentificatie en autorisatie zijn daarbij wel essentieel.

Uit tweets die tijdens en na de bijeenkomst rondgingen, proefde ik wat ongenoegen dat er nog geen concretere actie en stappen zichtbaar waren. Dat gevoel voor urgentie is begrijpelijk en lijkt niet onredelijk. De techniek om dit allemaal te realiseren bestaat gewoon al. Modulair, van datacontainers, via indexen van zoekmachines, tot interfaces van afzonderlijke portals. En de informatie is er eigenlijk ook al. Zowel de catalogusgegevens van al die deelnemende bibliotheken, als de metadata van het grootste deel van de tijdschriftartikelen.

Toch ligt daar nog even de bottleneck die snelle realisatie nog wat minder realistisch maakt. Want over al die informatie die in die infrastructuur moet worden aangeboden, zijn we niet zelf de baas. Voor al die catalogusrecords van OB's, UB's en andere deelnemende B's is dat eigenlijk een beetje gek, maar wel de realiteit. Voor metadata en inhoud van artikelen van al dan niet commerciële uitgevers is dat al veel begrijpelijker. Er zal dus nog heel wat moeten worden onderhandeld met allerlei partijen over rechten, tarieven, betalingen en licenties. Zoiets kan zelfs nog wel eens langer gaan duren dan een Nederlandse kabinetsformatie.

In de discussies op de bijeenkomst zelf kwam uiteraard nog ter sprake of we dit nu echt in Nederlands perspectief moeten oplossen. Zou dit niet meteen maar *Global* moeten worden? Tot de gewenste versnelling zal dat vast niet leiden, al zou de afkorting wel gewoon GII blijven.

Giga veel gegevens

Begin juli heb ik de leeftijd van 2 gigaseconde bereikt. Jammer genoeg had ik het toen te druk daar een feestje van te maken. Nu is dit ook geen gebruikelijke mijlpaal, waarvan vrienden je verwijten het niet gevierd te hebben. Maar het is wel een aardig moment om gevoel te krijgen hoeveel 2 miljard eigenlijk is. Voor een PC is 2 gigabyte intern geheugen erg krap bemeten. De mijlpaal van 2 miljard doorzoekbare webpagina's had Google acht jaar geleden al bereikt. Maar na 2 miljard seconde moet de mens al bijna met pensioen.

Begin juli werd in het kader van Linked Data initiatieven gemeld dat in de Verenigde Staten intussen meer dan 6 miljard computerleesbare overheidsgegevens online staan, in de vorm van zogenaamde RDF-tripels. Is dat veel? Dat lijkt inderdaad giga veel. Als je al die gegevens zou willen bekijken, daar een seconde per stuk voor uittrekt en dag en nacht zou doorgaan, dan heb je immers bijna drie mensenlevens nodig voor je ze allemaal gezien hebt.

Bij dit soort berichten (en aantallen) komt de vraag op hoe het eigenlijk met de Nederlandse overheid zit. Hoeveel gegevens heeft die online, liefst ook als Linked Data, als RDF-tripels? Daar lees je eigenlijk nauwelijks iets over. In elk geval bestaat de indruk dat we nog aanzienlijk achterlopen. Mederedacteur Jos van Dijk klaagde onlangs op IP's redactieblog al over de gebrekkige openbaarheid van Nederlandse overheidsinformatie.

Daarbij ging het Jos overigens ook om informatie die je niet zo makkelijk als "data" in RDF-tripels kunt weergeven. We mogen dus wel een voorbeeld nemen aan de VS. En aan Britse initiatieven, want ook daar is men erg actief op dit terrein.

In de VS lijkt het overigens niet eens zo zeer de overheid zelf te zijn die data als RDF beschikbaar stelt. Een heleboel is het resultaat van door de overheid gesteunde universitaire projecten, onder meer bij RPI, Renselaer Polytechnic Institute, de oudste technische universiteit van de VS in Troy. Bij mij roept die naam onverwachte herinneringen op, want in 1978 (1 gigaseconde geleden!) kwam ik regelmatig op de fiets door dat oude industriestadje ten noorden van New York.

Toch zijn aantallen niet alleenzaligmakend. Immers ook allerlei geografische en meteorologische gegevens behoren tot overheidsverzamelingen. En dan gaat het hard. Je hoeft maar gegevens van 4 meteorologische grootheden van 1000 plaatsen (voor de VS niet uitzonderlijk veel), op 24 tijdstippen per dag, gedurende 60 jaar verzameld te hebben en je zit al dik boven de 2 miljard gegevens. Veel hoeft dus niet altijd lekker te zijn. Het gaat ook om de aard van de gegevens die overheden publiekelijk beschikbaar maken. Maar dat is zeker geen argument om giga weinig te doen

Instantspagaat

Al geruime tijd bespeur ik een spagaat tussen de manier waarop Google het zoekers makkelijk probeert te maken en de manier waarop ik professioneel naar informatie wil zoeken. De overdreven nadruk dat het eerste resultaat het beste moet zijn, suggereert dat maar één antwoord goed kan zijn. Een mythe waar SEO-ers ijverig aan meewerken. Dat betekent dat webzoekmachines eigenlijk vooral optimaliseren voor wat informatiespecialisten van oudsher *known-item search* plegen te noemen. Vragen waarvan je vrijwel zeker weet dat er een antwoord is en eigenlijk ook maar één antwoord. Waar staat dat boek? Hoe laat vertrekt die trein? Waar koop ik een bed? Wat is de goedkoopste aanbieding voor die camera? Waar draait die film?

Zoekgoeroe Danny Sullivan schreef onlangs dat hij al bij Google had geklaagd, dat hun inspanning zich alleen nog richt op hun streven te zorgen dat de eerste hit het door de gebruiker gezochte item is. De verdere resultaten zijn eigenlijk helemaal niet meer zo goed, als je een gewone informatieve onderwerpsvraag stelt. Juist wat dat betreft is er een groot verschil tussen webzoekers als Google en software voor *enterprise search*. Die rankt ook, maar voor onderwerpsvragen in die omgeving blijven alle resultaten potentieel belangrijk. Dat verklaart waarom Exalead zo anders werkt dan de pure webzoekers.

De recente introductie van Google Instant versterkt die *known-item* nadruk nog extra. Naarstig probeert Google te voorspellen wat de bedoeling van je zoekvraag zal zijn, ook al heb je nog pas twee letters ingetikt. Een ultieme poging tot het vinden van de heilige graal van het doorgronden van onuitgesproken bedoelingen van zoekers, de *user intent*.

Maar die *intent* moet wel een beetje braaf zijn, want Phil Bradley's weblog signaleerde opmerkelijk gedrag bij woorden die zo aangevuld zouden kunnen worden dat ze te 'expliciet' materiaal opleveren. Probeer maar N-U-D: bij intikken van de D verdwijnen plotseling alle resultaten. Het lege scherm wordt pas weer gevuld als Google gezien heeft dat je als vierde zoekletter iets anders dan een E of I intikt. Waterdicht blijkt Google's kuisheidsgordel overigens niet; er blijven genoeg woordfragmenten die minder kuize plaatjes over je scherm laten flitsen.

Voor sommigen blijkt Google Instant intussen zeer afleidend of zelfs verslavend. Binnen enkele dagen werden al afkickpogingen gemeld. Onhandig is dat je Instant's speculaties wel met één muisklik kunt uitzetten, maar dat je het alleen ergens in je search-settings weer kunt inschakelen. Terwijl ResearchBuzz intussen ook al instant YouTube, Maps en Images meldde, sluit ik me meer aan bij Phil Bradley, die op Twitter verzuchtte: "The more I look at Google Instant, the more I hate it. Promotes poor, sloppy searching. Which is of course entirely to Google's benefit". Misschien niet lenig genoeg meer voor die spagaat tussen gemak en professionaliteit?

Bridging the indexer-gap

Ik denk dat Middelburg de Nederlandse stad is die het dichtst bij Engeland ligt. Wellicht heeft dat een rol gespeeld bij de Engelse Society of Indexers om vorige maand haar jaarlijkse congres niet in Engeland te houden, maar in Middelburg, hemelsbreed dichter bij Ramsgate dan bij Arnhem. De organisatie was overigens wel grotendeels in handen van Nederlandse indexers die lid van die vereniging zijn.

Indexers? Bij dat woord denk je als lezer van InformatieProfessional toch het eerst aan onderwerpsontsluiters zoals die bij bibliotheken werkzaam zijn. Al noemen we die meestal niet echt zo. De indexers waar het hier om ging, produceren dan ook heel andere producten dan verrijkte titelbeschrijvingen. Het zijn de professionele samenstellers van indexen die we achter in de betere non-fictie boeken aantreffen, de back-of-the-book-indexen.

Wellicht vraagt u zich af wat ik daarmee te maken had. Samen met twee collega's uit de eigen IP-kringen waren we daar drie van de sprekers. Dat had een paar bibliotheekcollega's trouwens op het verkeerde been gezet. Pas bij de eerste inleiding realiseerden ze zich op welk soort indexers dit congres gericht was. Natuurlijk zijn er wel inhoudelijke parallellen te bedenken tussen deze indexers en onderwerpsontsluiters. Beide moeten de inhoud van publicaties analyseren en bedenken met welke termen ze die toegankelijk maken. En dat op zo'n manier dat ze willekeurige gebruikers helpen zo effectief mogelijk bij de gewenste informatie terecht te komen. Alleen doen die indexers aan een extreme vorm van diepte-indexering. In hun indexen wordt een boek in feite op microniveau ontsloten.

Er was echter nog een andere parallel. Discussies van sommige aanwezige Engelsen - de wat oudere - veroorzaakten bij mij een sterk déjà-vu gevoel door de manier waarop men tegen de eigen beroepsbeoefening aankeek. Alleen liep die parallel niet erg synchroon in de tijd. De angst dat computers de professionaliteit van het vak onrecht aandoen en dat automatisering brodeloos zou maken, hebben we in de bibliotheekwereld toch wel al twintig jaar achter ons gelaten. Om nog maar niet te spreken over opgewonden discussies over al dan niet in bepaalde indextermen op te nemen punten, komma's en streepjes.

Dat het E-book deze indexers ook nieuwe kansen kan bieden tegenover de bedreiging van full-text doorzoekbaarheid van teksten, werd nog nauwelijks onderkend. Een door collega Evert Jagerman geëntameerde discussie dat digitale boeken nieuwe manieren van toegankelijkheid mogelijk maken, vond nog weinig weerklank. Toch kan een navigeerbare representatie van de inhoud en de kennis in een boek, via topic-map-achtige structuren, veel flexibeler en doorzichtiger toegang geven dan een ouderwetse boekindex. Maar misschien is dat juist wel te arbeidsintensief en is computerassistentie bij deze micro-ontsluiting essentieel om het te kunnen realiseren. Dat brengt die twee soorten indexers in de toekomst misschien dichter bij elkaar, nog dichter dan Middelburg en Engeland.

* A t*ranslation and adaptation of this column has been published in* ***The Indexer*** *(2011, vol. 29, nr. 2, p. 78)*
http://www.ingentaconnect.com/contentone/index/tiji/2011/00000029/00000002/art00009

Geen nieuws van Google

Een echt nieuwtje over Google zou zijn als eens een week niets nieuws over Google te melden zou zijn. Dat nieuwtje heb ik nog niet. Toch heb ik het wel weer eens over Google. In de Googoblogosphere en op Twitter verschenen in november berichten dat Google zijn zoekresultaten soms manipuleerde buiten reguliere ranking-parameters om. Door het effect van kleine wijzigingen in bepaalde zoekvragen te bekijken, kon men aantonen dat rankings voor bepaalde zoekvragen "hard gecodeerd" waren.

Dat hangt er wellicht ook mee samen dat Google lijkt te proberen een antwoordmachine te worden, in plaats van een zoekmachine. Een onderzoekje wees echter uit dat dat nog niet zo goed lukt. Ask, met de erfenis van AskJeeves, slaagt daar nog altijd beter in (ook beter dan Bing). Die transitie naar antwoordmachine geeft intussen vooral rumoer bij Search Engine Optimizers, die niet meer weten hoe sites optimaal te optimizen. Juist informatiespecialisten zouden zich echter zorgen moeten maken. Juist wij willen zelf in de hand hebben hoe onze zoekvragen worden geïnterpreteerd, in plaats van dat aan Google over te laten.

Hoe verhoudt dit zich tot de scherpe reactie van zoekgoeroe Danny Sullivan op geruchten dat de EU onderzoek doet naar bevoordeling van "eigen" informatie door Google, boven die van concurrenten? Sullivan vond dat de EU Google ten onrechte verweet zich als zoekmachine te gedragen; dat het nu eenmaal het prerogatief van een zoekmachine is om antwoorden te geven die aan een zoekvraag voldoen. Dat lijkt echter een wat erg gesimplificeerd beeld, gezien die gemanipuleerde antwoorden.

Een andere - tijdens de kerstvakantie al weer uitgedoofde - buzz was die over een Amerikaanse opticien die zulke slechte service gaf en zijn klanten zelfs bedreigde, dat alle fora overliepen van negatieve berichten. Maar ja, voor je PageRank telt elke link. Hoe ze over je praten doet er niet toe, als ze maar over je praten. Zo kwam de opticien steeds hoger in Google's rankings. Het lijkt wel op bezwaren tegen wetenschapsbeoordeling op basis van citatietellingen. Ook publicaties kunnen worden aangehaald omdat er zoveel mee mis is. Denk maar aan de koude kernfusie destijds. Met weblinks gaat die negatieve linkinflatie alleen nog veel harder.

Geschrokken door de opwinding heeft Google onmiddellijk zijn algoritmes aangepast. Bij voor PageRank uitgevoerde linktellingen wegen negatieve reacties nu veel minder (of zelfs negatief?). Heeft Google dan echt al zo goede automatische sentiment-analysis dat die betrouwbaar kan bepalen wat negatieve oordelen zijn? Ook nu waren de SEO-ers weer boos over deze ad-hoc ingreep in het ranking-mechanisme. Het laatste nieuwtje (met foto) was overigens dat die opticien in de boeien was geslagen, omdat hij het te bont maakte met het bedonderen van zijn klanten. Maar dat nieuws had niets met Google te maken.

Slow info

Slow Food is een trend waarmee we al vertrouwd zijn. Intussen was ik ook *Slow Scholarship* - langzame wetenschap - tegengekomen. Daar blijkt zelfs een heuse Facebook-groep met 59 leden voor te zijn. In onze jachtige twitterwereld leek *Slow Information* me daarom ook wel wat. Maar dat bleek ook al niet nieuw meer. Twee jaar geleden is al over een *slow information movement* geblogd. Zo'n beweging past mooi in het straatje van Nicholas Carr. Tijdens mijn januari-vakantie was ik hem weer tegengekomen in een lang artikel in El Pais. Aanleiding was de Spaanse vertaling van zijn boek "The Shallows". Zijn daarin uitgesponnen ideeën dat de vluchtige informatieovervloed en de haastig te googelen weetjes op internet zijn diepere denk- en concentratievermogen aantasten, had hij drie jaar geleden al geuit in een essay in The Atlantic Monthly.

In mijn column van december 2008 was ik daarop al eens ingegaan. In diezelfde column mompelde ik ook nog wat over tegenstrijdigheid van op internet te vinden informatie over het al dan niet slimmer worden door frequent zoekmachinegebruik. Dat ging dus over betrouwbaarheid van informatie, iets waaraan *slow information* wel eens een bijdrage zou kunnen leveren. Zo waren uitspraken van Google CEO Eric Schmidt over onze - zeker niet slowe - informatieproductie afgelopen najaar al te snel overal op internet gerepliceerd. Schmidt beweerde dat we nu elke twee dagen zo'n 5 exabyte aan informatie produceren, net zo veel als in de hele periode vanaf het begin der tijden tot 2003.

Pas recent kwamen er langzaamaan reacties van mensen die achterliggende cijfers hadden gezocht en daaraan gerekend hadden. Die kwamen erop uit dat Schmidt zeker een factor 20 fout zat; 40 dagen klinkt dan toch wat minder indrukwekkend dan twee. In dit verband kan ik nog op andere oude columns terugkomen. In 2007 (column 27) - en zelfs al in 2001 (column 12) - betoogde ik dat je informatie niet in bytes moet meten. Een video van een lezing van 200 MB bevat niet wezenlijk meer informatie dan de uitgeschreven tekst van 20 kB. Die cijfers van Schmidt moeten we dus met nog heel andere ogen bekijken. De vorm waarin informatie geregistreerd is, kan ook nog factoren 10.000 verschil maken.

Vergeleken met blogs en Twitter is deze column ook maar *slow info*. Niet alleen omdat ik er lang(zaam) over zit te broeden, voordat ik een definitieve versie opstuur. Maar vooral omdat het productieproces van een papieren blad nog extra vertraging oplevert voordat het op uw mat ligt. Daardoor bent u op het moment van lezen misschien al te laat om Nicholas Carr live over zijn boek te horen praten. Met *fast info* had u tijdig kunnen weten dat hij 2 maart in Amsterdam bij het John Adams Institute spreekt. Daarvoor was ik wel snel genoeg.

* Nicholas Carr / Is Google Making Us Stupid? What the Internet is doing to our brains - https://www.theatlantic.com/magazine/archive/2008/07/is-google-making-us-stupid/306868/

* Our Cluttered Minds (New York Times Sunday Book Review 2010) - https://www.nytimes.com/2010/06/06/books/review/Lehrer-t.html

Carr's gelijk

Nicholas Carr heeft kennelijk toch gelijk. Zijn laatste boek, The Shallows, heb ik nog steeds niet gelezen, maar ik laat me intussen wel afleiden door alle berichtjes die ik via RSS en Twitter binnenkrijg. Ook door Carr's eigen weblog Rough Type.

Maar laat ik bij het begin beginnen. Zoals vorige maand aangekondigd, ben ik inderdaad naar Carr's lezing in Amsterdam geweest. Daar bleek wel enige nuance te zitten in het impliciet positieve antwoord op zijn suggestieve vraag van drie jaar geleden: *"Is Google making us stupid?"*. In feite beargumenteerde hij nu dat niet Google ons dom maakt, maar SMS, E-mail, Twitter, Facebook en smartphones. Die zorgen voor een voortdurende stroom aan afleiding. Zo zijn smartphones van hulpje tot despoot geworden.

De zo ontstane "interruptierijke omgeving" heeft ons doen wennen aan een voortdurend bombardement van informatie, dat ons afleidt van activiteiten die eigenlijk serieuze concentratie vergen. Neurowetenschappers hebben al lang aangetoond dat ook volwassen hersens nog zeer plooibaar zijn. Gewenning aan afleiding maakt dus dat onze hersens ook steeds meer afgeleid willen worden.

Kunnen we daar nog van afkicken? Desgevraagd zei Carr dat alleen de elite zich nog kan veroorloven af te haken. Voortdurend online zijn is al zo mainstream dat de gewone man zich dat al niet meer kan permitteren - hoezo *digital divide*? En moeten we dat wel willen?

Carr ziet ook wel pluspunten aan de nieuwe media. Naast het fantastische gemak om aan allerlei informatie te komen, verbetert ook onze visuele en ruimtelijke intelligentie. Maar weegt dat op tegen de minkant dat het ten koste gaat van ons analytisch, zorgvuldig, kritisch denkvermogen. Daarvoor is langer volgehouden concentratie nodig dan moderne media ons toestaan.

Intussen kwam in Carr's weblog van 7 maart een ander aspect aan de orde: de informatie-overload. Uiteraard duikt Clay Shirky op, met zijn uitspraak van enkele jaren geleden: *"information overload is filter failure"*. Carr onderscheidt echter twee soorten overload, *situation overload* en *ambient overload*. Die *situation overload* is het probleem dat met goede filters is op te lossen, dus de overload waar Shirky het over had. Maar dat probleem is volgens Carr opgelost. Daar hebben de eerder versmade Google's van deze wereld intussen voor gezorgd.

Maar wat die niet hebben opgelost - en ook niet kunnen - is de *ambient overload*. Dat is de echte overvloed, de overdaad die overblijft nadat alles wat niet relevant is, al is weggefilterd. Er is echt teveel informatie. Zinnig, relevant, interessant, lezenswaardig. Ook als we ons niet laten afleiden door berichtjes op onze smartphones, blijft er geen tijd over om dat allemaal te lezen. Misschien moeten we wel concluderen dat dat de paradox is: juist doordat we steeds beter kunnen filteren, kunnen we die relevante informatie ook allemaal te zien krijgen. Die vormt zelf al de afleiding, inclusief Carr's eigen stukkies.

* Rough Type - http://www.roughtype.com/

* Information Overload is Filter Failure, Says Shirky - https://lifehacker.com/5052851/information-overload-is-filter-failure-says-shirky

De paradox van de ontsluiting

Bibliothecarissen en informatieprofessionals zijn waarschijnlijk de enigen die bij het woord *ontsluiting* denken aan het soort ontsluiting waaraan bibliothecarissen en informatieprofessionals denken als ze het woord ontsluiting horen. Maar ik krijg wel eens de indruk dat zelfs zij binnenkort eerder aan barensweeën dan aan trefwoorden zullen denken. Dat komt dan niet (alleen) vanwege de plannen tot beëindiging van de Gemeenschappelijke OnderwerpsOntsluiting.

Het is een veel algemener verschijnsel. Iedereen verwacht dat computers intussen volledig in staat zijn om van alle informatie de betekenis te doorgronden. Dat die foutloos kunnen zoeken en ons op elk moment in elke situatie precies van de juiste relevante informatie kunnen voorzien. Dat zoekmachines inderdaad steeds beter met sommige digitale tekstuele informatie om kunnen gaan wordt moeiteloos doorgetrokken naar alle andere soorten informatie. Ontsluiting en ontsluitingssystemen zijn daarmee achterhaald; ze zijn ouderwets en onnodig arbeidsintensief.

Goed, dan laten we het verder aan computers over. Die kunnen het wel zonder ons af. Bovendien komt het semantisch web er aan. Web 3.0 zal in één klap alle problemen oplossen die we misschien toch nog een enkele keer hebben met het vinden van relevante informatie. In het kader van die verwachting is het aardig om eens achter de schermen te kijken van dat semantisch web en van de Linked Data die daarvoor gebruikt worden.

Daar blijkt het over te lopen van een overvloed aan metadata, metadatasystemen, metadatastandaarden en ontologieën. Veel kennisorganisatiesystemen die met die modieuze benaming "ontologie" gelabeld worden, zijn in feite gewoon ontsluitingssystemen.

We moeten dus concluderen dat dat semantisch web alleen kan bestaan dankzij de aanwezigheid van ontsluitingssystemen - in de breedste zin - en dankzij het feit dat zo veel informatie wordt ontsloten. De bouwers van dat semantisch web, de nerds die alles weten over RDF, SKOS, OWL en nog veel meer alfabetsoep, die prijzen ons bibliothecarissen de hemel in, dat wij zo goed zijn in het ontsluiten van informatie, dat we daar zulke mooie systemen voor hebben bedacht, dat we daarmee al zoveel materiaal hebben ontsloten. Alleen daardoor kunnen de bouwers van het semantisch web verder werken aan de semantiek en interoperabiliteit van "ons soort informatie". Wij hoeven dus niet meer te ontsluiten, omdat het semantisch web dat overbodig maakt. En dat semantisch web kan dat overbodig maken, doordat wij alles zo goed ontsluiten (of op zijn minst ontsloten hebben). Voelt u de paradox?

Wat intussen wel achterhaald is, is het idee dat we nog altijd alles zelf zouden moeten ontsluiten. Dat is het mooie van dat semantisch web, dat iedereen gebruik zal kunnen maken van wat anderen al gedaan hebben, maar voor een deel dus ook nog zullen moeten blijven doen. Want helemaal zonder ons kunnen computers het toch ook nog niet af.

2000 Wikipedia's

Recent moest ik wat schrijven over tagging en folksonomies. Daarbij komt de vraag boven wat mensen ertoe beweegt om materiaal op internet van tags te voorzien - het dus te ontsluiten. Die activiteit is goed te verklaren als daar eigenbelang bij is - het zelf weer kunnen terugvinden van je eigen materiaal, je foto's op Flickr, je boeken in LibraryThing.of je bookmarks bij Delicious (dat gelukkig recent uit handen van Yahoo! gered is). Maar waarom zou je helpen materiaal van anderen te ontsluiten, zoals onbekend fotomateriaal dat archieven op internet zetten of videomateriaal van Beeld en Geluid, ook al gebeurt dat laatste via een spelelement zoals in hun project Waisda.

In dat verband kwam ik ook al een aantal keren de naam van Clay Shirky tegen, dezelfde die ik hier eerder aanhaalde naar aanleiding van zijn uitspraak "information overload is filter failure". Nu ging het om zijn vorig jaar uitgekomen boek *Cognitive Surplus: Creativity and Generosity in a Connected Age*. Als ik besprekingen en verwijzingen mag geloven, legt hij in dat boek veel meer de nadruk op de positieve kanten van ons huidige internettijdperk dan bijvoorbeeld Nicholas Carr. Volgens Shirky hebben veel mensen behoefte iets zinnigs te doen met (een deel van) de vrije tijd die ze hebben. Dat daar veel van is blijkt uit cijfers dat Amerikanen 200 miljard uur per jaar voor de TV doorbrengen.

De New York Times bracht dat al tot de opmerking "For a vast majority of us, watching TV is essentially a part-time job". Shirky rekent voor dat die tijd genoeg zou zijn om 2000 Wikipedia's vol te schrijven.

Shirky maakt duidelijk dat hij het mooi en zinnig vindt dat mensen dat "intellectuele overschot" uit de titel van zijn boek inzetten voor het toevoegen van kennis aan internet. Daarmee staat hij wel lijnrecht tegenover somberman Andrew Keen die in zijn boek *The cult of the amateur* aangaf geen al te hoge dunk te hebben van het niveau van dergelijke kennis. De paar keer dat ik Keen in Nederland die opvatting persoonlijk heb zien (en horen) uitdragen, deed hij me trouwens een beetje denken - ook qua uiterlijk - aan de Zwarte Zwadderneel, een onheilsprofeet uit de Bommelstrips, die ook bij zonnig weer altijd nog een zeer lokaal regenbuitje boven zijn hoofd heeft hangen. Hoewel Keen erg ver gaat met zijn hetze tegen alle *user-generated content*, of dat nu weblogs zijn of de Wikipedia zelf, ben ik toch ook wel blij dat die 2000 Wikipedia's per jaar niet worden volgeschreven. De *filter failure* zou dan wel eens al te groot kunnen worden. Maar dat een heel klein percentage van Shirky's "cognitive surplus" ook kan helpen om allerlei kennis door middel van tags te ontsluiten, is wel mooi meegenomen.

Too inconvenient to go after it

"If it is too inconvenient, I'm not going after it": Convenience as a critical factor in information seeking behaviors. Dat is de pakkende titel van een binnenkort in *Library and Information Science Research* te verschijnen artikel. De titel sprak mij ook wel aan. Want bij de opzet van Omega, het Utrechtse zoeksysteem voor digitale wetenschappelijke publicaties, was - al weer bijna tien jaar geleden - één van de uitgangspunten dat het er niet ingewikkelder mocht uitzien dan Google. Ook toen voorzagen we al dat het anders niet (genoeg) gebruikt zou worden.

Datzelfde uitgangspunt lijkt ook ten grondslag te liggen aan de "Discovery Tools" als Summon, Primo en Meresco, die de laatste tijd in gebruik komen bij steeds meer Universiteitsbibliotheken. Dat zijn de nieuwe zoeksystemen voor een combinatie van de digitale tijdschriftartikelen en de catalogus (en tevens de vervanging van de oude zoekfunctie daarvan). Is voor die *discovery tools* trouwens al een acceptabele Nederlandse term bedacht? "Ontdekkingsgereedschap" klinkt niet echt lekker.

Recent - nog niet zeer diepgaand - onderzoek naar zoekgedrag van wetenschappers in Utrecht, onthulde dat - met uitzondering van PubMed - de bibliografische vakdatabases niet erg veel meer als informatiebron worden gebruikt. Zelfs een mooie database als PsycInfo niet. Dat zou ook wel eens kunnen komen doordat veel van de bijbehorende zoeksystemen "*too inconvenient*" worden gevonden.

Zeker in vergelijking met een wetenschappelijke zoekmachine als Google Scholar - wat voor terechte kritiek je ook moge hebben op de kwaliteit daarvan.

In elk geval sluit die conclusie mooi aan bij een ervaring van mijn collega Jeroen Bosman. Nadat hij in zijn college informatievaardigheden aan een groep Utrechtse studenten het zoekinterface van Ovid had uitgelegd, reageerde één van de studenten met de woorden: "Leuk hoor, maar u denkt toch niet dat ik dat ga gebruiken". Zelf heb ik onlangs in de VOGIN-cursus weer een groep echte informatieprofessionals wegwijs gemaakt in datzelfde zoeksysteem. Dat heeft er wel toe bijgedragen dat ik me wat beter in die reactie van Jeroens student kan verplaatsen.

Het is dat de inhoud van de betreffende databases soms ook in *discovery tools* geïntegreerd wordt. Anders zou ik moeten concluderen dat aanbieders van bibliografische zoeksystemen de databaseproducenten in feite de nek omdraaien, omdat ze het zoeken in die databases *too inconvenient* maken, In tijden van informatieovervloed zijn zoekers dan niet meer bereid dergelijke systemen te gebruiken *to go after it*. In tijden van krimpend budget zullen bibliotheken vervolgens hun abonnementen op die weinig gebruikte databases opzeggen, hoe professioneel ze ook zijn samengesteld en hoe perfect ze ook mogen zijn ontsloten. Voor databaseproducenten lijkt die waarheid me ook heel *inconvenient* te zijn.

* “If It Is Too Inconvenient, I’m Not Going After It:” Convenience as a Critical Factor in Information-seeking Behaviors (Library and Information Science Research, 2011) - https://www.oclc.org/content/dam/research/publications/library/2011/connaway-lisr.pdf

Filosofie van de komkommer

Komkommers waren dit voorjaar al zoveel in het nieuws geweest, dat er afgelopen vakantieperiode geen klassieke komkommertijd meer afkon. De afwisseling van Murdochschandaal, Breiviktragedie, Winehousedrama en Kredietplafondcrisis zorgde dat er genoeg echt nieuws was om onderweg in Engeland dagelijks een krant te kopen - ondanks het feit dat wij daar wel goed weer hadden. Neveneffect was dat ik in The Guardian een interessant stukje tegenkwam over hyperlinkstructuren in de Wikipedia. Informatieprofessionals blijven ook in hun vakantie alert.

Het beschreef een experiment waarbij je begint bij een willekeurig Wikipedia-lemma en vervolgens telkens de eerste inhoudelijke hyperlink naar een ander Wikipedia-lemma aanklikt. Na een beperkt aantal klikken - meestal tussen de 10 en 30 - blijk je in meer dan 94% van de gevallen bij het lemma "Philosophy" als eindpunt uit te komen, hoe triviaal het beginpunt ook was. Een betere aanwijzing dat de filosofie de moeder van al ons weten is, kun je nauwelijks bedenken. Toen ik dit thuisgekomen verder nazocht, bleek in de Wikipedia zelf ook al een soort spelletje aan dit verschijnsel gewijd te zijn.

Maar in het krantenartikel kwam een nog opmerkelijker uitkomst aan de orde. Het verschijnsel beperkt zich tot vooral de Engelstalige Wikipedia. In de Nederlandse treedt het vrijwel nooit op! Daar kom je hoogstens in eindeloze loops terecht, wat een Nederlandse editor deed verzuchten dat Nederlanders kennelijk alleen in cirkeltjes redeneren. Toch is het een interessant probleem wat de oorzaak van dit opmerkelijke verschil zou kunnen zijn. Vast niet dat Engelsen en Amerikanen zo veel filosofischer zijn ingesteld dan Nederlanders

De verklaring die Engelstalige Wikipedianen hadden bedacht waarom zij bijna altijd bij filosofie uitkomen, kan ons misschien wat verder brengen. Zij merkten op dat de links die je bij dit spelletje achtereenvolgens aanklikt, je geleidelijk "omhoog" brengen, naar steeds algemener onderwerpscategorieën, zoals in de hiërarchie van een classificatie. En de filosofie, als koningin der wetenschappen, staat eenzaam aan de absolute top daarvan.

Maar waarom dan niet in het Nederlands? Mijn hypothese is dat Nederlanders niet in classificaties denken, omdat ze daar niet mee zijn opgegroeid. Waar Engelsen of Amerikanen bij elk decimaal getal meteen de bijbehorende onderwerpscategorie uit Dewey of Library of Congress classificatie weten te noemen, hebben wij daar niets mee. Doordat we geen Dewey kennen, linken we ook niet hiërarchisch. Dat is meteen een mooie verklaring voor onze nationale karaktertrek waarom we zo anti-autoritair zijn en ook in het maatschappelijk verkeer geen hiërarchie erkennen. Of is het net andersom, dat wij door die karaktereigenschap een hiërarchische SISO of Dewey verafschuwen? Wie kan deze hypotheses eens falsifiëren?

Hoe dan ook: in de Nederlandse Wikipedia zul je van het lemma "komkommertijd" (het bestaat echt) dus nooit bij filosofie terechtkomen. Iets dat in deze column wel gelukt is.

* The Only Way Is Essex + Wikipedia = philosophy; The online encyclopedia's unlikely routes to enlightenment (The Guardian, 2011) - https://www.theguardian.com/technology/2011/jul/10/only-way-essex-wikipedia-philosophy

---

Noot: Exact acht jaar geleden (september 2003 – column 20) bleek ik het ook al eens over komkommers en de komkommertijd gehad te hebben

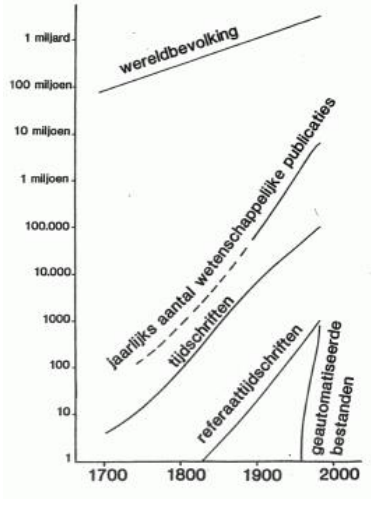
Publish, but perish anyway

Het zelf verzinnen van onderzoeksgegevens door de Tilburgse sociaalpsycholoog Diederik Stapel wordt hier en daar wel verklaard als uitvloeisel van de prestatiedruk die onderzoekers voelen. "Publish or Perish" en bovendien niet zo maar iets publiceren, maar beslist iets baanbrekends. Of deze psychologische verklaring voor databedrog juist is of niet laat ik hier even in het midden - psychologen vertrouw ik per definitie niet meer. Wel onderstreept dit schandaal nog eens het belang van datarepositories voor het kunnen controleren van onderzoeksresultaten. Data doen er inderdaad wel degelijk toe.

Nog even terug naar dat "Publish or perish". Bij mijn eerste colleges over online literatuur zoeken, 30 jaar geleden, had ik een mooie grafiek gemaakt van de geschatte groei van de aantallen wetenschappelijke artikelen en wetenschappelijke tijdschriften. Dat plaatje was grotendeels gebaseerd op cijfers van Derek de Solla Price, de goeroe van de wetenschap van de wetenschap. Het jaarlijks aantal wetenschappelijke artikelen bleek in de 230 jaar tussen 1750 en 1980 met een factor 100.000 te zijn toegenomen. Over die hele periode toonde de grafiek een bijna voortdurende exponentiële groei, met een verdubbeling iedere 14 jaar. Zoiets is uiterst ongebruikelijk. In de natuur duren periodes van exponentiële groei altijd aanzienlijk korter. Lang voordat 230 jaar om zijn of een verschijnsel met vijf ordes van grootte gegroeid is, treedt altijd verzadiging op, doordat er tekort ontstaat aan voedsel, grondstof of ruimte. Daardoor krijg je een soort afvlakkende S-curve.

Voor de aantallen publicaties verwachtte ik dat eigenlijk ook. Zelfs het publish-or-perish-effect zou daar volgens mij niet tegenop kunnen. Op een gegeven moment is de groei van het aantal wetenschappers op en kunnen ze niet nog meer onderzoekjes per dag doen. Omdat ik te lui was elk jaar echte nieuwe cijfers op te zoeken, tekende ik dat grafiekje elk volgend collegejaar gewoon weer ietsje verder door in de tijd, maar wel al een beetje afvlakkend, de S-curve in. Begin dit jaar kwam ik erachter dat ik daarmee de zaak aardig besjoemeld had. Want de NRC van 12 maart meldde dat ook de laatste decennia, de 30 jaar sinds 1980, de wetenschappelijke productie nog altijd elke 12 jaar verdubbelt. Kennelijk had ik buiten Chinezen en dikke duimen gerekend. Met mijn grafiekjes begin jaren '90 was ik dus geen haar beter geweest dan Diederik Stapel. Alleen haalde ik de kranten niet met opzienbarende onderzoeksresultaten - en dus ook niet met datagesjoemel.

Dit alles doet me sterk denken aan een cartoon uit The New Yorker die ik in mijn college ook altijd op de overhead-projector legde. Op een kerkhof kijkt een heer neerslachtig naar een grafschrift: "*Russell P. Dinsmore / 1907-1984 / Published, but perished anyway*".
In overdrachtelijk zin geldt dat laatste zeker ook voor Diederik A. Stapel.



Onvoltooid online verleden

Via Linked-in kreeg ik een paar weken geleden een berichtje van oud-vakgenoot Jak Boumans, waarin hij meldde dat een boek van hem was uitgekomen. "Toen digitale media nog nieuw waren - Pre-internet in de polder (1967-1997)". Dat heb je met een generatie die wat ouder wordt, dacht ik toen, die heeft de neiging in het verleden te duiken. Maar verder vergat ik het meteen weer. Tot ik zondagochtend ineens Jak's stem uit de radio hoorde komen. In OVT, VPRO's geschiedenisprogramma over de Onvoltooid Verleden Tijd.

Het ging over zijn boek. Precies verslag doen van dat vraaggesprek kan ik niet, want op zondagochtend ben ik nog niet zo'n geconcentreerd luisteraar. Wel hoorde ik plotseling een oud radiofragment waarin de zachte G van Chriet Titulaer langs kwam. Voor ons - onkundige luisteraars - introduceerde de goeroe van de techniek van de toekomst een nieuwe online dienst van de PTT: Viditel. In augustus 1980 door staatssecretaris Neelie Kroes gelanceerd. Voor onze jonge lezertjes: Viditel zag eruit als Teletext, met net zo weinig tekst als nu nog altijd op ons TV-scherm verschijnt. Maar dit kwam via een modem over de telefoonlijn binnen. En daar kon je informatie opvragen! Als ik Chriet's woorden goed onthouden heb, kostte een abonnement toen 1700 gulden per jaar. **[*]**

Nog afgezien van de prijs - zelfs zonder inflatie meer dan we nu voor breedband internet betalen - vond ik Viditel toen al niks. Echte informatieprofessionals met hun geavanceerde IBM pc's, waar wel drie keer zo veel lettertjes op het groene scherm pasten, konden al een veelheid aan serieuzere databases raadplegen bij allerlei online hosts. Waarbij ik gemakshalve vergat hoeveel duurder dat nog weer kon worden, met betaling voor elke gebruikte minuut. Maar inderdaad, Viditel is nooit doorgebroken. De opmerking van Jak Boumans dat Chriet Titulaer Nederland rijp gemaakt heeft voor het digitale tijdperk, kwam mij dan ook overdreven voor. Pas toen we internet kregen en dat goedkoop en makkelijk genoeg werd, heeft Nederland die digitale wereld echt omhelsd. En dan vooral dankzij een generatie die Chriet nauwelijks meer bewust heeft meegemaakt.

Terugkijkend kun je trouwens zeggen dat die online informatie toen eigenlijk al "in the cloud" zat, of het nu van Viditel, Dialog of Dimdi kwam. Alleen moest die term nog worden uitgevonden. Een groot verschil is natuurlijk de digitale informatieschaarste van toen tegenover de overvloed van nu. In tijden van schaarste ben je bereid te betalen en moeite te doen met ingewikkelde systemen. In tijden van overvloed moet alles vanzelf gaan en moet een dienst zelfs een "belevenis" zijn om gebruikt te worden. Ter illustratie deze vacature van de toekomst die ik net zag langskomen: *The University of Michigan Library is seeking an innovative, and talented user experience professional to join our User Experience Department.*

***** *Jak Boumans liet mij achteraf weten dat dit de prijs voor de apparatuur was; het abonnement bleek veel goedkoper.*

Het jaar van de webscale discovery

2011 lijkt het jaar van de "webscale discovery systems" te zijn geworden. Althans voor veel grote bibliotheken in het hoger onderwijs. De heilige graal van de contentintegratie lijkt allerwege gerealiseerd te kunnen worden. Het is niet meer nodig lokaal de content te verzamelen waarop men licenties heeft en die door een eigen zoekmachine te laten indexeren, zoals bij een voorloper als het Utrechtse Omega-systeem gebeurt. Alles kan aan externe dienstenleveranciers en hun "cloud" worden uitbesteed. De helft van de UB's en een aantal HBO bibliotheken heeft intussen zijn keuze gemaakt voor één van de vier leveranciers die in Europa de dienst uitmaken.

Het is interessant eens te kijken naar de toekomst van dergelijke geïntegreerde zoeksystemen in de context van collecties. Of eigenlijk in de context dat je steeds minder van echte collecties kunt spreken. Voor tijdschriften was dat in feite al zo. Door de *big deal* hebben universitaire bibliotheken al bijna identieke tijdschriftcollecties. Namelijk alles wat de uitgevers uitgeven die in de consortiumcontracten zitten. Maar ook voor boeken lijkt het die kant op te gaan.

Het vorige nummer van IP berichtte over lezingen door Rick Anderson over *Patron Driven Acquisition*. Simplificerend komt dat er op neer, dat in principe alle boeken (voor zover digitaal) voor de klant beschikbaar zijn, en dat op een boek dat een klant echt begint te lezen, automatisch een licentie wordt genomen - en ook pas op dat moment.

Maar dat betekent dat die klant dus elk willekeurig boek in een zoeksysteem (nee, in HET geïntegreerde zoeksysteem) moet kunnen vinden. En ook dat alle zo beschikbare boeken effectief eigenlijk al tot de collectie behoren - en tot ieders collectie.

Als er, op wat klein lokaal spul na, toch geen eigen collecties meer zijn, zou je de conclusie kunnen trekken dat je wel toekan met Google-Scholar + Worldcat als zoeksysteem en dat je geen dure discovery service meer nodig hebt. Wel blijft dan het probleem van de "delivery" van de informatie die je ontdekt hebt. Als geen collectie meer gedefinieerd is, hoe organiseer je dan de toegang. Bij PDA kan ik me voorstellen dat dat goed gaat, al zijn er vast nog wat organisatorische en financiële probleempjes op te lossen. Maar bij tijdschriftartikelen moet misschien nog echt wat gebeuren. Ook al vierden we onlangs Open Access week, het is voorlopig nog een utopie te denken dat alles voor iedereen gratis toegankelijk wordt.

Voor tijdschriften stelde Rick Anderson, als opvolger van de "big deal", een "tiny deal" voor, waarbij bibliotheken alleen nog maar betalen voor artikelen die echt door klanten gelezen/gedownload worden. Tarieven daarvoor moeten gemiddeld dan wel optellen tot wat bibliotheken nu ook ongeveer kwijt zijn. Alleen dan kun je zowel uitgevers als bibliotheken enigszins tevreden houden. Maar of 2012 al het jaar van zo'n nieuwe *webscale delivery* kan worden, valt nog te bezien.

Too much to know

Volgens James Gleick, auteur van "The Information; a history, a theory, a flood", was 1948 het jaar waarin het informatietijdperk begon. Dat jaar werd de transistor uitgevonden, formuleerde Shannon zijn informatietheorie en introduceerde die ook het woord "bit". Daarmee was de basis gelegd voor dat informatietijdperk. Zelf behoor ik nog net tot de laatste generatie die daarvoor geboren is - een soort stenentijdperkmens dus. Toch ben ik altijd druk met informatie bezig geweest. Veertig jaar geleden nog gewoon op papier, toen ik stapels afleveringen van gedrukte Physics Abstracts doorploegde, op zoek naar achtergrondliteratuur voor mijn promotieonderzoek naar *roosterfouten in silicium*.

Veertig jaar later houd ik me vooral nog bezig met informatie om de informatie. Er lijkt nu ook veel meer van te zijn. Iedereen klaagt over information overload. Maar is dat echt alleen een probleem van nu? In een blogpost begin januari werd die vraag beantwoord met de uitroep "There Has Always Been Too Much To Know!". Daarbij werd gerefereerd aan de hier al eens door mij genoemde "slow internet movement" (wat niet slaat op de snelheid van uw dataverbinding). In die blogpost werd ook Gleick's boek aangehaald: de filosoof en wiskundige Gottfried Leibniz sprak in 1680 al van een "horrible mass of books which keeps on growing" en de dichter Alexander Pope in 1728 van een "deluge of authors".

Dat de jaarlijkse productie aan wetenschappelijke publicaties in 300 jaar bijna een miljoen keer zo groot is geworden, lijkt niettemin bewijs van ongehoorde overvloed. Een verdubbeling elke 14 jaar heeft dat veroorzaakt. Dat betekent dat er dit jaar toch maar 7x zoveel bijkomt als 40 jaar geleden. Niet echt om van te schrikken, al is het meer dan mijn salaris in die tijd gegroeid is. Maar hoe zit het dan met de data-tsunami waarover we steeds lezen? Dat we elk jaar al 2x zoveel bytes produceren als het jaar daarvoor. Dat zijn maar bytes en geen informatie (al weet ik niet of Shannon het met die redenering eens zou zijn geweest).

In die 40 jaar is wel veranderd dat, dankzij Google en Webscale Discovery Tools, alles nu binnen enkele milliseconden vindbaar is, zonder tussenkomst van slimme informatiespecialisten. Maar werken we daardoor ook efficiënter? Besteden we minder tijd aan vergaren en selecteren van informatie? Aan vergaren waarschijnlijk wel. Of worden we ook hebberiger en willen we net als vrouwtje Piggelmee steeds meer? Dan moeten we weer strenger wieden en selecteren. Maar of we ons werk sneller en vooral beter zijn gaan doen dan 40 jaar geleden, weet ik niet. Ik moet dan altijd denken aan de vorig jaar overleden psycholoog Wagenaar die ons voorhield dat we vanzelf roekelozer gaan rijden als verplichte veiligheidsgordels ons een gevoel van extra veiligheid geven. Met ons informatiegebruik gebeurt vast net zo iets.

Google valt op zijn Face

Het einde van Google is al heel wat keren aangekondigd. Soms als *wishful thinking*, omdat Google voor sommigen de rol van *evil empire* van Microsoft heeft overgenomen. Soms vanwege veranderend informatiegedrag, omdat men liever door zijn sociale cirkels op Facebook geïnformeerd wil worden dan zelf te moeten zoeken. Soms uit ontevredenheid met Google als dienst - vooral door informatieprofessionals, die hun eigen zoekbehoeften als maatgevend zien. Tot die laatste categorie behoor ik zelf. Zelfs professioneel voelen we ons zo afhankelijk van Google, dat elke verslechtering bijna als persoonlijk affront wordt gevoeld, zeker als Google het beter denkt te weten dan wij zelf.

Is er dan zoveel mis met Google? Heeft u even?

- Google introduceert met veel poeha diensten waarop niemand zit te wachten en die dus ook weer snel verdwijnen (Buzz, Wave, ...).
- Google beëindigt diensten die mensen wel leuk vinden (Desktop search, Google Toolbar, Wonderwheel, Timeline, Code search, Google Square en nog 40 andere ...).
- Google verandert elke week de functionaliteit op zijn zoekscherm.
- Google verdonkeremaant ineens de link naar het geavanceerde zoekscherm, eerst een beetje door het onder een tandwieltje te verbergen, dan helemaal zolang je nog niks gezocht hebt, en stopt het dan toch weer standaard onder dat tandwieltje.
- Google laat de link naar zijn vertaalhulp verdwijnen, eerst onder het tandwieltje, dan eventjes helemaal en dan toch weer onder het tandwieltje.
- Google verstopt de "related" en "cache" links in de site-preview pop-up.
- Google denkt beter te weten wat ik bedoel dan ik zelf en past voortaan zonder toestemming te vragen mijn zoekvraag aan.
- Google gaat ongevraagd ook zoeken op allerlei enigszins gelijkende of verwante woorden waarin ik helemaal niet geïnteresseerd ben.
- Google vindt het ineens niet meer nodig dat alle ingetikte zoektermen in het zoekresultaat voorkomen en bepaalt eigenmachtig welke mogen ontbreken.
- Google schaft de + syntax voor exact zoeken af, omdat de nieuwe dienst Google+ een centrale rol is toebedacht.
- Geschrokken van alle negatieve reacties op bovenstaand gedoe introduceert Google ijlings de "Verbatim" functie waarmee je wel letterlijk (of eigenlijk "woordelijk") zoekt op wat je intikt.
- Google voert ongevraagde personalisatie van zoekresultaten via sociale cirkels verder door om Google+ sterker als concurrent van Facebook en Twitter te promoten.
- ... en dan heb ik vast nog een heleboel vergeten.

Maar wat gebruik ik zelf als ik iets zoek? Nog steeds Google! Arghhhh.

Geloof ik nu echt dat Facebook Google zal verdringen? Op ReadWriteWeb blog zag ik net een oproep om je massaal bij Facebook af te melden, omdat Facebook met zijn problematische privacy en kapitalistische beursgang de rol van *evil empire* al weer van Google lijkt over te nemen. Misschien gaat Facebook dus ook wel op zijn gezicht.

Op Facebook's divan

Stephen Wolfram, de man achter de Wolfram|Alpha feitenzoekmachine, blogde begin maart dat hij zijn leven had laten analyseren. Niet op de divan bij zijn psychiater, maar door een zelfgeschreven computerprogramma. Dat was een combinatie van de door hem ontwikkelde "Mathematica" software en de analysemogelijkheden van de betaalde versie van Wolfram|Alpha. Dat destilleerde "zijn leven" uit de ruim 300.000 mailtjes die hij sinds 1989 had verstuurd. Een flink deel daarvan aan zichzelf - soms automatisch gegenereerd - over wat hij dagelijks deed. En ook uit de ruim 1,5 miljoen mailtjes die hij in die tijd van anderen had ontvangen en uit de meer dan 100 miljoen toetsaanslagen die hij had gedaan (en geregistreerd - je bent een data-junk of niet). En ook was statistiek losgelaten op vergaderingen en telefoontjes. En zelfs op het aantal stappen dat hij had gemaakt op een tredmolen die hij gebruikte om tijdens zijn werk nog wat lichaamsbeweging te krijgen. Zo zou ik nog wel even door kunnen gaan.

Maar deze eigen analyse van een persoonlijk leven wil ik eigenlijk afzetten tegen de bezorgdheid over onze privacy bij webdiensten, waar je helemaal niet zelf in de hand hebt wat er allemaal geregistreerd en geanalyseerd wordt. Denk bijvoorbeeld aan de analyses die Facebook kan maken op basis van de gegevens die het over gebruikers verzamelt. Dat dat meer is dan je denkt, bleek uit enkele anecdotes in een lezing van Anne Helmond op het Bobcatsss-congres eerder dit jaar.

Een Oostenrijkse student die wilde weten wat Facebook over hem wist, ontving na veel getouwtrek uiteindelijk een document van meer dan 1200 bladzijden, uitgeprint een bijna 20 cm dikke stapel A4tjes, het resultaat van drie jaar Facebook-gebruik. En dat scheen nog lang geen compleet overzicht te zijn.

Dat er zo veel is, komt voor een deel doordat Facebook's "like"-button, op welke webpagina die ook staat, automatisch gegevens over het browsegedrag van gebruikers terugmeldt, oorspronkelijk zelfs zonder dat die knop beroerd werd. Hoewel dat intussen verboden is, blijven er nog altijd onvoorstelbare hoeveelheden data binnenkomen. En niet alleen bij Facebook, maar minstens zoveel bij Google en Twitter. Zie maar in Helmond's bijdrage aan het recente Unlike-us congres.

Misschien moeten we dus toch maar onze toevlucht nemen tot de *Web 2.0 suicide machine*, waarmee je digitale zelfmoord kunt plegen, naar keuze op Facebook, Myspace, Twitter of LinkedIn. Gegarandeerd *"no resurrection possible"*. Hoewel een digitaal kunstproject, kwam het zo serieus over, dat de bedenkers aanvragen van uitvaartondernemingen ontvingen, die deze optie in hun dienstenpakket wilden opnemen. Blijft de vraag of we de analyse van ons leven aan de Facebooks van de digitale wereld willen overlaten of het liever zelf in de hand willen houden, via tools als Wolfram|Alpha of heel radicaal door digitaal suicide te plegen.

* Anne Helmond / Visualizing Facebook'salternative fabric of the web (2012) - http://www.annehelmond.nl/2012/03/12/visualizing-facebooks-alternative-fabric-of-the-web/

Het ondiepe web

Ineens stond het diepe web weer in de belangstelling. Merkwaardig genoeg kwam dat door de AIVD. In een rapport over het gevaar van Jihadisten werd gesuggereerd dat hun gekonkel zich op het diepe web afspeelde. En de AIVD moest natuurlijk even uitleggen wat dat was: *"Het zichtbare deel van het internet, ook wel oppervlakteweb genoemd, is dat deel van het internet dat wel terugvindbaar en doorzoekbaar is. Wetenschappers schatten in dat het onzichtbare web 550 keer groter is dan het zichtbare web. Dit wil zeggen dat het onzichtbare web meer dan 99,8% van het totale web uitmaakt en dat dus minder dan 0,2% van het web tot het zichtbare deel behoort."*

Harde cijfers uit een onverdachte bron zou je denken. In de praktijk viel dat anders uit. NRC Next's factchecking afdeling wierp zich onmiddellijk al op deze bewering. Terecht, want dat cijfer 550 was gebaseerd op een onderzoek van BrightPlanet (Lyman & Varian) uit 2000. Destijds was dat in dit blad in de www.wie.wat.waar-rubriek ook al aan de orde geweest. Behalve NRC's harde conclusie "onwaar" is daar nog wel meer over te zeggen.

Elders in dit nummer vindt u mijn schatting dat de omvang van dat zogenaamde oppervlakteweb sinds 2000 met minstens een factor 1000 is toegenomen. De AIVD heeft dus óf niet doorgehad dat het web de laatste 12 jaar gegroeid is, óf ze denkt dat dat diepe web - al die databases van NASA, meteo's, LexisNexis en dergelijke - ook wel met diezelfde factor gegroeid zal zijn.

Zou dat echt? Bovendien weten onze speurneuzen kennelijk evenmin dat zoekmachines intussen succesvolle pogingen ondernemen om dat diepe web ook te indexeren, al dan niet in samenwerking met eigenaren van die databases. Vermoedelijk is dat diepe web sindsdien dus ook nog aanzienlijk minder diep geworden.

Maar het belangrijkste bezwaar is misschien wel dat helemaal niet duidelijk is wat met die factor 550 eigenlijk wordt bedoeld. Overigens kleefde dat bezwaar natuurlijk al aan dat oorspronkelijke onderzoek van BrightPlanet uit 2000. Wat probeerden ze te meten? In welke eenheden is gemeten? Meet je items? En wat zijn dan items? De middagtemperatuur op 23 april 1934 op Ameland uit een KNMI-database, is dat een item dat net zo hard meetelt als een PDF-document van een omvangrijk AIVD-rapport (voorzover dat niet geheim is)? En tellen al die steeds veranderende dynamisch gegenereerde oppervlaktewebpagina's elk apart mee? Of meet je mega/giga/exabytes, waardoor dat dikke PDF-document juist weer onevenredig veel meer credit krijgt dan die temperatuur op Ameland?

Kortom, het was onzin allemaal. En ik blijf met een akelig gevoel achter. Als een uitspraak van de AIVD over iets waarvan ik verstand denk te hebben, zo onzinnig blijkt, hoe betrouwbaar zijn dan hun (ook nog geheim gehouden) uitspraken over al die andere onderwerpen. Hun onderzoeken gaan kennelijk niet erg diep.

* Lyman & Varian / How Much Information? 2000 - http://groups.ischool.berkeley.edu/archive/how-much-info/

Bubble boys

Ik ben nog uit de tijd dat informatiespecialisten hun stinkende best moesten doen om met uitgekiende zoekstrategieën zo volledig mogelijk alle relevante informatie boven tafel te krijgen. In de huidige tijd van informatieovervloed lijkt dat steeds meer een anachronisme. Iedereen wil eigenlijk minder vinden, maar dat mindere moet dan wel precies zijn toegesneden op persoonlijke behoeften. Google hinkt daarbij een beetje op twee gedachten. Enerzijds proberen ze ons steeds meer te laten vinden door ongevraagd ook te zoeken naar varianten, samenstellingen en synoniemen van onze zoekwoorden. (Overigens lijken ze dat weer wat bij te regelen, nadat ze daarmee wel erg ver waren doorgeschoten). Anderzijds doen ze verwoede pogingen zoekresultaten te personaliseren op basis van wat ze al van je weten. En dan kom je in Eli Pariser's "filter bubble" terecht.

Wie herinnert zich nog die oude Seinfeld-aflevering met de "bubble boy", een immuundeficiënt jongetje dat in een plastic "bubble" leeft om niet met levensbedreigende virussen en bacteriën in aanraking te komen. Ook na 1992 is die aflevering nog eindeloos op de Nederlandse TV herhaald. Pariser vreest dat we met supergepersonaliseerde informatievoorziening in net zo'n plastic cocon terechtkomen, waardoor we niet meer met onwelgevallige meningen in aanraking komen. Wie neemt de rol van eigenwijze George Costanza op zich, om onze bubbles te laten ontploffen, maar dan zonder dramatisch gevolg? Zoekmachine DuckDuckGo beweert dat te doen. Geen personalisatie en ook meteen geen privacy-problemen, omdat ze geen informatie over je zoekgedrag bewaren. Het bracht weblog SearchEngineLand er al toe om DDG te afficheren als "The Biggest Long-Term Threat To Google".

Een heel andere manier om uit onze bubble te ontsnappen biedt zoekmachine MillionShort, onlangs in Phil Bradley's weblog gemeld. Die laat resultaten van de hoogstgeklasseerde 1.000.000 (of 100.000 of 10.000) websites weg uit zijn antwoordlijsten. Dan krijg je eindelijk ook eens resultaten te zien die niet door duur betaalde SEO-experts omhooggepompt zijn.

Een probaat middel om beslist niet alles te vinden is het zoeken in tweets. Er is geen zoekhulpmiddel dat verder terug dan de laatste twee weken de hele tweetosphere doorzoekt. Met SnapBird kun je wel onbeperkt terug in de tijd, maar dan moet je je zoekactie beperken tot de tijdlijn van één persoon, bijvoorbeeld die van jezelf. En dan zit je natuurlijk weer in een extreem nauwe bubble. Nog een ander nieuwtje (geloof ik) is Twicsy, de Twitter Pics Engine. Nog sterker dan bij gewone webplaatjeszoekers, zijn zoekingangen om twitterfotoos te vinden natuurlijk beperkt tot alleen de paar woorden uit de bijbehorende tweets. Maar daardoor juist leuk om uit te proberen en je te laten verrassen door totaal onverwachte vondsten en associaties. Twicsy breekt dus juist uit de bubble. En voor de bubble *boys* onder u: "expliciete" content lijkt niet gefilterd te worden.

Dingen of documenten

Al een tijdje laten zoekmachines zich erop voorstaan eigenlijk geen zoekmachine meer te zijn. Dat ze niet alleen zoeken, maar ook zelf beslissen welke informatie de gebruiker wil weten. Met Wolfram|Alpha was dat heel duidelijk. Maar ook Bing afficheerde zich bij zijn introductie als "decision engine" in plaats van als zoekmachine. Overigens was dat niet erg terecht, want Bing leverde gewoon net zulke lijstjes resultaten als Google en alle andere zoekmachines. En nu is Google zelf dan ook zo ver met zijn Knowledge Graph, in Googles eigen Search Blog, "Inside Search", aangeprezen met de kreet: "Things not strings". Dat houdt in dat Google zich niet meer beperkt tot de domme reeksen letters in onze zoekvragen, maar weet heeft van een heleboel dingen of entiteiten, waartussen het relaties kent. En waarover het ook allerlei gegevens en feiten beschikbaar heeft, op grond waarvan dubbelzinnigheden ontrafeld kunnen worden.

Hoe komt Google eigenlijk aan al die gegevens? Op de discussielijst van Freebase, een grote open source semantische database, sprak Google al zijn dank uit aan alle mensen die geholpen hebben Freebase te vullen. En dat die wel blij mogen zijn dat hun werk dankzij Google veel ruimere verspreiding krijgt: "This is bringing a lot of the concepts about high-quality, structured data and entity disambiguation, that Freebase has refined, to a much wider audience."

Behalve die Freebase-gegevens, aangevuld met Wikipedia en CIA World Factbook, beweert Google nog gebruik te maken van eigen gegevens, waardoor er meer dan 500 miljoen "dingen" in die Knowledge Graph zouden zitten en meer dan 3,5 miljard feiten en relaties. En bovendien: "... it's tuned based on what people search for, and what we find out on the web." Maar waarschijnlijk dan toch nog vooral voor de Engelstalige wereld.

Toch vraag ik me sterk af of alle soorten gebruikers hierop zitten te wachten. Wie alleen even een feitje wil opzoeken is er vast mee geholpen, maar had dat natuurlijk ook allang in de Wikipedia kunnen vinden. Het is echter de vraag of dit voor de echte informatieprofessional wel zo nuttig is. In veel gevallen wil die liever gewoon een lijstje documenten die hij zelf op waarde beoordeelt en waaruit hij zelf conclusies, voors en tegens, alternatieven en dergelijke destilleert. Of liever gezegd, dat hoort een goede informatieprofessional te willen. Zelf aan het stuur, en het niet aan Google overlaten.

Is het dan helemaal mis? Nee, gelukkig blijft Google ook gewoon met lijstjes resultaten komen; documenten in plaats van dingen. En het is wel een volgende stap in de richting van het semantisch web. Ook al noemt Google dat zelf niet zo en vermijdt het zelfs - tot verdriet van sommige semantisch web enthousiasten - ook maar enig jargon uit die hoek te gebruiken bij het aanprijzen van deze nieuwe functionaliteit.

Puriteinen of Piraten

Informatie is meestal geen centraal thema waarop politieke partijen elkaar bij de verkiezingen te lijf gaan. Maar voor het eerst doet nu een partij mee, waarvan het hoofdbeginsel wel om informatie draait: de Piratenpartij is "voor een vrije informatiesamenleving". Eindelijk weten dus alle informatieprofessionals waarop ze zonder verdere bedenkingen kunnen stemmen. Want een vrije informatiesamenleving, daar moeten we allemaal wel voor zijn. Maar ja, eigenlijk ben ik ook wel "voor de dieren". Hoe moet ik dat nu tegen elkaar afwegen? Dat krijg je met die single issue partijen. Nou zijn dergelijke partijen in de praktijk meestal lang niet zo single issue als hun hoofdbeginsel of hun naam doet vermoeden. Zelfs de Partij voor de Toekomst zal zich misschien nog wel een beetje met het verleden bezig houden. Maar wat zeggen die piraten nu echt allemaal over de informatiesamenleving?

Op de fact-checking website van NRC Next, next.checkt, is al een drietal stellingen van die partij tegen het licht gehouden. Twee daarvan hadden inderdaad wel wat met informatie te maken. De eerste was dat dit jaar zo'n 1,1 miljard wordt uitgegeven voor 10 miljoen aan medicijnen. Dat wil zeggen dat die medicijnen niet meer dan dat hadden hoeven kosten, als we maar wat anders met patenten zouden omgaan. De argumenten waarmee NRC's fact-checkers die stelling als onwaar afwezen, zijn door de piraten op hun beurt al weer afdoende onderuit gehaald. Maar met opvattingen over de wetgeving rond patenten had die discussie niets te maken.

Een andere stelling die was onderzocht had ook al met patenten te maken. De bewering dat in Nederland patenten zonder controle op nieuwheid worden toegekend, dat dus vrijwel alles gepatenteerd kan worden en dat je toekenning alleen achteraf bij de rechter kunt aanvechten, werd door NRC wel als waar beoordeeld.

Je zou hieruit misschien concluderen dat patenten de enige soort informatie zijn waarin de Piratenpartij geïnteresseerd is. Dat is zeker niet het geval. Dit was de toevallige selectie van de NRC fact-checkers. Bij de zeven kernpunten uit het partijprogramma van de Piratenpartij gaat het ook over privacy en over betere bescherming van persoonlijke en vertrouwelijke gegevens, over een meer open overheid, over hervorming van zowel het auteursrecht, als het merkenrecht, als (inderdaad) het patentensysteem en over de burgerrechten in die eerder genoemde informatiesamenleving. En onder dat laatste vallen dan onder meer het recht op toegang tot internet, op vrije communicatie en op vrijheid van meningsvorming en meningsuiting.

Op 12 september dus allemaal op de piraten stemmen? Misschien goed om voor die tijd nog even op bibliofuture.nl te kijken. Niet echt een kieswijzer, maar daar worden wel alle partijprogramma's doorgelicht op wat ze over Openbare Bibliotheken zeggen. Misschien levert dat een uitweg voor wie te puriteins is om op woeste piraten te stemmen.

Feiten en fabels

Vorige maand had ik het al even over fact-checking. Dat had niet te maken met de NOS-baas die deze zomer beweerde dat "feiten toch niet bestaan". Heftige reacties dwongen hem toen tot nuancering, maar ik geloof niet dat het journaal intussen fact-checkers in huis heeft. Nee, mijn opmerking vorige maand hield verband met partijprogramma's en verkiezingen. Verkiezingen waarvan u intussen allang de uitslag kent, maar ik bij het schrijven van deze column nog niet. Ik moet het ergens op een Grieks eilandje zien te achterhalen; maar dergelijke fact-finding zal ook daar niet moeilijk meer zijn.

Verkiezingen geven kennelijk extra aanleiding tot fact-checking. Niet alleen hier, maar ook in de VS. Misschien dat het daar nog wel meer nodig is dan hier, want wat daar in verkiezingstijd voor onzin wordt uitgekraamd, zowel over als door kandidaten, is meer dan een zinnig mens kan weerleggen. Toch schijnt daar een soort revolutie plaats te vinden, waarbij de pers de taak van waarheidsvinding gelukkig weer op zich neemt. Men spreekt in dit verband al van "post-truth politics". In *The Atlantic* stond 29 augustus bijvoorbeeld een stuk onder de titel "Bit by Bit It Takes Shape: Media Evolution for the 'Post-Truth' Age". En de site van *Poynter* (een instituut voor journalistiek) vroeg zich een dag later naar aanleiding van de Republikeinse conventie af "Did the media just enter age of 'post-truth politics' with Paul Ryan speech?"

Nu bestond in Amerika in bepaalde kringen altijd al een cultuur om feiten te controleren. De site van Factcheck.org staat voortdurend vol met claims die als "false" worden afgedaan. Naar voorbeeld daarvan was de Tilburgse journalistiekopleiding in 2008 ook begonnen zijn studenten allerlei beweringen te laten uitzoeken, onder de naam FHJ-Factcheck. Hoewel ik er de laatste tijd niets meer over gehoord had, blijkt fhjfactcheck.wordpress.com buiten de hogeschoolvakantie nog altijd aardig actief. Toch werd fact-checking pas echt mainstream toen ook hier de kranten zich ermee gingen bemoeien. Eerst gratis krant De Pers met zijn Pinokkio rubriek. Hoe langer de neus, hoe minder waar. En na opheffing daarvan nam NRC Next dat idee over met "Next.checkt". En sinds kort is er nu ook www.factcheck.nl en treden checkers zelfs bij TV-debatten op.

Maar moeten we dit alleen aan de journalistiek overlaten? Fact-checking is een kerncompetentie van informatiespecialisten (waar onderzoeksjournalisten trouwens ook toe behoren) . Die kunnen goed zoeken, kunnen bronnen op betrouwbaarheid beoordelen en kunnen zo feiten van fabels scheiden. Laten IP-ers zich daar dus ook op profileren, want goed zoeken is nog altijd een vak. Of, zoals onze Belgische collega Rosemie onlangs twitterde: "*Ondanks alle tools blijvend nood aan searchers: methodisch, sterk gestructureerde detectives en analisten*". Dat is precies wat je voor fact-checking nodig hebt. En eigenlijk mag je fact-checkers dan zelfs wel REsearchers noemen.

* Bit by Bit It Takes Shape: Media Evolution for the 'Post-Truth' Age (The Atlantic, 2012) - https://www.theatlantic.com/politics/archive/2012/08/bit-by-bit-it-takes-shape-media-evolution-for-the-post-truth-age/261741/
* Did the media just enter age of 'post-truth politics' with Paul Ryan speech? (Poynter, 2012) - https://www.poynter.org/news/did-media-just-enter-age-post-truth-politics-paul-ryan-speech

Over archieven en petabytes

Hoeveel een Petabyte is, zou u waarschijnlijk niet geweten hebben als het Internet Archive niet net gevierd had intussen 10 Petabytes te hebben opgeslagen. Voor wie dat nieuws gemist heeft: een petabyte is een miljoen gigabyte. En voor wie zich geen voorstelling kan maken van dat soort getallen: 10 Petabyte is evenveel bytes als er seconden verlopen zijn in alle levens samen van van alle nu levende Nederlandse vrouwen (of alle mannen natuurlijk). Overigens omvat dat Internet Archive nog heel wat meer dan alleen de bij veel informatieprofessionals bekende WaybackMachine. Of echte archivarissen dat ook allemaal "archief" zouden noemen, weet ik niet.

"Het" archief was afgelopen maand overigens toch al in het nieuws, nu de fusie van KB en Nationaal Archief ineens (voorlopig?) niet blijkt door te gaan. Hoewel we al heel wat jaren verhalen horen dat bibliotheek en archief naar elkaar toegroeien, blijf ik in de praktijk nog altijd een flinke kloof zien. En dat betreft niet alleen de wettelijke aspecten die als voornaamste reden voor het uitstel van de KB-NA-fusie werden aangevoerd. Het gaat ook verder dan de verschillende manieren waarop archivarissen en bibliothecarissen aankijken tegen ontsluiten en toegankelijk maken van hun materiaal. Weblog "De bronnen van Clio" analyseerde in dat verband wat er mis ging in andere landen die wel fuseerden.

Zelf proef ik altijd nog aanzienlijk verschil in algehele cultuur. Eerder dit jaar was ik bijvoorbeeld bij een symposium over het als Open Data beschikbaar stellen van erfgoedmetadata. Museum en bibliotheek zagen daar niet zoveel problemen om dat onder CC-0 te doen, een Creative Commons licentie waarbij je zelfs niet eist dat gebruikers vertellen dat de gegevens oorspronkelijk bij jou vandaan komen. Maar een spreker uit de archiefhoek zag dat nog helemaal niet zitten.

Toch komt dat misschien nog goed. In het kader van "Hack de Overheid", houdt Opencultuurdata.nl een hackaton waarbij iedereen wordt uitgenodigd Apps en andere toepassingen te ontwikkelen, op basis van (meta)data die vanuit de erfgoedsector beschikbaar gesteld worden. Van de 35 datasets die nu op de lijst staan, blijken er vijf afkomstig van het Nationaal Archief en nog eens zes van regionale archieven. Maar drie van de 35 worden expliciet geafficheerd als afkomstig van een bibliotheek, waarvan twee van de KB.

Anderzijds is terughoudendheid van archiefzijde wel te begrijpen. Zij hebben immers veel privacy-gevoelig materiaal. En men wil voorkomen dat nog eens hele ladingen persoonlijke gegevens zomaar ongecontroleerd ergens anders terecht komen. Zoals al die persoonsgegevens die de mormonen in het verleden hebben weten los te praten, om die onderin hun berg in Utah te digitaliseren. Om vervolgens al die ongelovige - althans niet-mormoonse - Nederlanders alsnog postuum te kunnen dopen, zodat ze voor het mormoonse hiernamaals gered worden. Hoeveel petabyte zouden die mormonen daar intussen hebben opgeslagen?

* Geen fusie Koninklijke Bibliotheek en Nationaal Archief (De bronnen van Clio, 2012) - http://hethistorischatelier.blogspot.com/2012/10/geen-fusie-koninklijke-bibliotheek-en.html

Sterke merken gaan teloor

Veertig jaar geleden was ik druk met mijn onderzoek aan roosterfouten in silicium. Rond de meetopstelling stonden nogal wat kastjes en apparaten van Hewlett Packard. Dure, maar zeer betrouwbare spanningsbronnen, hoogfrequent oscillatoren en oscilloscopen. HP had toen een internationale, innovatieve, maar qua omvang wat beperkte markt. De afgelopen jaren stond op mijn bureau, op mijn beide werkplekken, weer een kastje van HP, nu PC genaamd. Misschien iets minder betrouwbaar, maar in elk geval aanzienlijk goedkoper dan welk van die vroegere kastjes ook. En daarmee een onvergelijkbaar veel grotere markt bedienend. Maar nog altijd hardware, echte apparaten.

Tien jaar geleden waren we in Utrecht op zoek naar nieuwe zoekmachinesoftware voor Omega, het Webscale Discovery systeem dat we gebouwd hadden (al noemde niemand dat toen nog zo). Daarvoor kwamen we terecht bij Autonomy. Dure maar zeer betrouwbare software. Want - volgens de folders - gebruikmakend van een combinatie van Bayesiaanse waarschijnlijkheidsrekening én de informatietheorie van Shannon. Dat moest wel goed zijn. Bovendien bij allerlei grote bedrijven in gebruik voor enterprise search. Een week na onze beslissing Autonomy te kiezen, werd bovendien bekend dat ze de andere marktleider, het Amerikaanse Verity, hadden overgenomen. De oprichter en directeur van Autonomy, Mike Lynch werd dan ook het paradepaardje van de Britse innovatieve industrie. Voor IP reden zich door Autonomy's PR-afdeling te laten verleiden om Mike Lynch, op doorreis via Schiphol, door mij te laten interviewen (december 2006).

Eén jaar geleden was een nieuwe HP-directeur ineens niet meer tevreden met zijn positionering als apparatenbouwer. Software leek lucratiever dan hardware. En dan vooral enterprise software. Tot veler verbazing werd voor $ 10 miljard Autonomy overgenomen. Tot Brits verdriet dat zijn paradepaardje kwijt raakte, maar Mike Lynch kon nu zeker gaan rentenieren. Als Autonomy-gebruiker kreeg ik in Utrecht ineens mailtjes van HP in plaats van de luxe uitnodigingen van Autonomy. Maar levert "betrouwbaar x betrouwbaar" weer "betrouwbaar" op?

Dat het zakenleven geen wiskunde is, bleek enkele weken geleden, toen plotseling bekend werd dat HP een recordverlies had geleden en dat dat door Autonomy kwam. ReadWriteWeb blog kopte "Autonomy's Irregularities Give HP An $8.8B Headache". Want zaken rond Autonomy schenen bij de verkoop veel rooskleuriger te zijn voorgesteld dan ze werkelijk waren. Wederom uit RWW's blog: "*... that some former members of Autonomy's management team used accounting improprieties, misrepresentations and disclosure failures to inflate the underlying financial metrics of the company, prior to Autonomy's acquisition by HP*".

Oeps, dat klonk fors. Had ik zes jaar geleden op Schiphol eigenlijk met een zwendelaar zitten praten? Had ik me zowel bij de UB Utrecht als bij IP door mooie marketingpraatjes laten inpakken? Voorlopig denk ik dat Autonomy's oude klanten zich vooral zorgen maken hoe het verder gaat met een product waar ze tevreden over waren; met die technologie was niets mis.

* Autonomy-baas Mike Lynch: Verity was een goede acquisitie (IP, 2006) - http://sieverts.pbworks.com/f/ip-autonomy.pdf

Infoclutter

"I've never really read newspapers or magazines, as I dislike the fact that most of it goes unread and they clutter my space". Dat zijn niet mijn woorden. Ik las ze in een reactie onder een bericht over de laatste papieren aflevering van Newsweek die eind December uitkwam. Maar die woorden spraken me wel aan, omdat ik thuis net mijn volledig verclutterde werkplek van overbodig papier had ontdaan. Een stapel van meer dan een meter hoog had ik net afgevoerd naar de papiercontainer. En dat waren dan nog niet eens kranten en tijdschriften, maar knipsels, printjes en copieën van artikelen waarvan ik ooit gedacht had, dat ze interessant genoeg waren om ze nog eens echt te lezen, of er iets mee te doen voor een onderzoekje of een column. Nee dus.

Wat mijn vakmatige interesse wekt, blijken vooral practische en toepassingsgerichte artikelen. Maar hoe practischer en toepassingsgerichter de beschreven onderwerpen, hoe sneller die artikelen al weer achterhaald en verouderd zijn. Het is nauwelijks de moeite waard ze langer dan een paar maanden te bewaren. En zeker geen vijftien jaar, zoals ik aan de data in mijn stapel zag. Het was nu eenmaal geen wetenschap waar het over ging. En zelfs bij echte wetenschap zit veel dat maar een korte halfwaardetijd heeft en dat onze databases en repositories vercluttert.

Wat te denken van een recent bericht dat meldde dat de Way-back-machine intussen 240 miljard URL's geregistreerd heeft? Die webpagina's worden met enige regelmaat gearchiveerd. Ook daar zitten veel tijdgebonden practische berichtjes bij, met een heel korte halfwaardetijd. Moet dat echt allemaal ongeselecteerd bewaard blijven. Ook al wordt mijn werkplek daar niet fysiek door verclutterd, toch vraag ik me wel eens af wat onze nazaten ooit nog eens moeten met al die opgeslagen petabytes.

In dit verband ook nog iets over Facebook's nieuwe zoekmachine. Daarvan wordt gemeld dat hij niet het web doorzoekt, maar alleen Facebook's "graph". Een netwerk van gegevens van een miljard mensen en 240 miljard foto's, waartussen een biljoen - in het Engels *trillion* - verbindingen bestaan, als ik Facebook mag geloven. Of daar veel feitelijks tussen zit waarin ik als professioneel zoeker geïnteresseerd ben, betwijfel ik. Anders dan Henk van Ess, heb ik nog geen toegang, zodat mijn oordeel berust op voorbeeldvragen die ik in allerlei berichten langs zie komen - maar die zijn niet van het soort dat ik pleeg te stellen. Bovendien lijkt de beantwoordbaarheid sterk afhankelijk van de hoeveelheid privé-informatie die je zelf al op Facebook gedeeld hebt. ReadWriteWeb blog bleek nog niets opwindends aan die zoekmachine te zien en zelfs te vinden dat hij vrije gebruikers weer opsluit binnen een middelmatig en onvolledig product. Toch moet ik binnenkort maar eens kijken of de nieuw bedachte slimme filtermogelijkheden iets kunnen met alle infoclutter aan trivialiteiten die Facebook biedt.

Informatie moeilijk gemaakt

In een aflevering van Folia Magazine, het universiteits-/hogeschoolblad van HVA + UVA, stond onlangs een interview met een natuurkundetweeling. Allebei hoogleraar, de een in de VS, de ander bij de UVA. Gezamenlijk - zoals tweelingen betaamt - hadden zij nieuwe theoretische inzichten ontwikkeld rond de zwaartekracht. Dat zou geen fundamentele kracht zijn (hoezo Higgs?), maar een gevolg van de informatiehuishouding van de ruimte. "De ruimte is een opslagplaats van informatie over waar deeltjes zich bevinden."

Als deeltjes zich verplaatsen verandert die informatie. En dat is de zwaartekracht. Dat klinkt bijna als het woordgebruik dat je op sommige alternatieve pseudomedische websites tegenkomt. Maar dit zijn natuurkundehoogleraren en geen kwakzalvers (mag ik hopen). Het is wel pijnlijk dat zelfs een tot informatiespecialist omgeschoolde natuurkundige zoals ik, geen idee heeft op wat voor manier informatie tot natuurkundige wetten over zwaartekrachtvelden zou kunnen leiden. En wat bedoelen ze hier eigenlijk met "informatie"? Bedoelen ze echt gegevens over de posities van deeltjes? De coördinaten van alle atomen in het hele universum? Dat is nog eens big data.

Overigens is dit niet het enige voorbeeld van het gebruik van het woord informatie dat mijn pet te boven gaat.

In eerdere columns (o.a. januari 2012) heb ik wel eens nonchalant de informatietheorie van Shannon laten vallen, alsof ik daar alles van afwist. Nou vergeet het maar. Die theorie gaat over signaalverwerking, datacompressie, datacommunicatie en datacodering. Met zo min mogelijk bits zorgen dat toch geen informatie verloren gaat. Het begrip entropie wordt daar ook bijgehaald, een begrip dat ik alleen kende uit - ook weer - de natuurkunde, uit de thermodynamica. Shannon schijnt onder meer te hebben afgeleid dat "*de kortst mogelijke manier om berichten in een bepaald alfabet te coderen, gelijk is aan hun entropie gedeeld door de logaritme van het aantal symbolen in het betreffende alfabet*". Bent u er nog? Shannon publiceerde zijn baanbrekende artikel over de informatietheorie precies 65 jaar geleden, maar ik begrijp dat die theorie nog heel populair is en allerminst met pensioen gaat. Het is wel een heleboel wiskunde. Vroeger schrok ik daar nooit zo van. Misschien moet ik me er toch maar eens in gaan verdiepen, nu ik zelf wel met pensioen ben.

In allebei die voorbeelden gaat het over heel andere zaken dan waar wij, eenvoudige informatiespecialisten aan denken, als we het over informatie hebben. Misschien wel jammer dat diepere gedachten over wat informatie is, bij ons zo ongeveer ophouden bij de vaak als pyramide getekende opeenvolging data>>informatie>>kennis. In dat schema zouden wij de informatie waarover Shannon en die zwaartekrachttweeling het hebben, misschien ook liever data noemen. Kennelijk denkt de een bij het woord informatie dus aan heel wat anders dan de ander. Zou men dat ooit nog eens goed in een ontologie kunnen krijgen, zodat het semantisch web wel precies begrijpt wat al die verschillende mensen ermee bedoelen?

* C.E. Shannon / A Mathematical Theory of Communication (1948) - http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Hoe digitaal is mijn bibliotheek

In 1995 noemden we het nog een "Elektronische Bibliotheek". Althans in Utrecht, waar de UB toen een project onder die naam afsloot. Dat moest de opstap worden tot wat nu een digitale bibliotheek heet. En is die er ook gekomen? Ja natuurlijk, die is er al lang. En niet alleen in Utrecht. Daarom was ik nogal verrast hoeveel commotie ontstond over het plan van Hogeschool InHolland om op een digitale bibliotheek over te gaan. Wat me daarbij vooral verbaast is de richtingenstrijd die dat anno 2013 - 18 jaar na die Utrechtse Elektronische Bibliotheek - nog blijkt op te leveren.

Aan de ene kant degenen die de aankondiging van Doekle Terpstra met de grootste afkeuring ontvingen. Aangevoerd nog wel door de NRC, die zonder wetenschappelijke onderbouwing beweerde dat wie van een beeldscherm moet lezen, onvermijdelijk studieachterstand oploopt.

Aan de andere kant de "believers" die zondermeer beweren dat alles wat je nodig hebt over twee jaar digitaal kan zijn, ongeacht de aard van het materiaal.

Sommigen lijken zelfs te denken dat alles al digitaal is, als het maar in een digitale catalogus zit. Over die catalogus herhaal ik nog maar eens wat ik al tien jaar verkondig, dat je daar in huidige vorm helemaal niets aan hebt. Althans als je hem voor onderwerpsvragen wilt gebruiken. Een opvatting die pas nu gemeengoed lijkt te worden.

Voor dat onderwerpszoeken zijn de huidige Discovery systemen een veel beter hulpmiddel. Het woord "discovery" zit daar niet voor niets in. Wel inconsequent om daar dan ook je oude catalogus in te willen stoppen, want die was nu juist ongeschikt voor discovery.

Zo'n Discovery systeem is natuurlijk een mooie opstap naar een digitale bibliotheek, zoals InHolland wil. Maar je moet dan wel genoeg digitale abonnementen hebben, die de discovery-leverancier ook nog in zijn systeem moet kunnen opnemen. Dat kan voor een Hogeschool - meer dan voor een universiteit - nog wel een knelpunt vormen.

Voor mijn eigen lessen bij de HvA - ook een hogeschool - placht ik mijn studenten vrijwel altijd digitale literatuur aan te bieden. Gewoon blogposts, websites en PDF's. Daartoe werd ik niet gedwongen door een college van bestuur of door een bibliotheek (die in mijn gebouw trouwens niet fysiek aanwezig was). Dat besloot ik gewoon zelf, als eenvoudig docent, bottom up. Waarom? Omdat het zoveel makkelijker was. Omdat ik makkelijker nieuwe dingen kon opgeven. Omdat ik veel fijnmaziger kon bepalen wat mijn studenten moesten lezen. En als er eens een echt boek nodig was, moesten ze dat gewoon kopen. Wat ik niet weet - en achteraf misschien te weinig heb onderzocht - is hoeveel mijn studenten daarvan ook echt gelezen hebben. En ik heb al helemaal geen dubbelblind onderzoek gedaan of mijn digitale leesvoer studievertraging heeft opgeleverd.

PS: *De waarschuwing op de UvA-site "Digitale Bibliotheek verdwijnt per 1 maart 2013" klinkt in dit kader wel wat verrassend.*

Weg met de Apps

Onlangs zag ik via Twitter een attendering op een nieuwe aflevering van de Informatie Curator. Leuk, want *curation* (liever in het Engels want dan geeft het wat minder connotatie met faillissementen) is helemaal in; een oude taak van de informatieprofessional in een nieuw jasje. Maar helaas kwam ik niet ver, want ik kon er alleen bij met een App voor een iPad. Het kon niet in mijn gewone browser op mijn gewone PC, waar ik al mijn andere informatiewerk mee doe. Dat kom ik vaker tegen, dat voor tamelijk gewone dingen - zoals zo'n curation tool - verplicht een aparte App nodig is.

Dat geeft een sterk déjà-vu gevoel. Terug naar vroeger, toen je op je PC voor allerlei dingen aparte programmaatjes nodig had. Terug naar vroeger, toen je voor zoeken naar informatie in tien verschillende systemen moest zoeken. Toen de informatie die je nodig had nog over verschillende incompatibele informatiesilo's verspreid zat. Toen we nog niet gehoord hadden van content integratie en van Discovery tools. Maar intussen kennen we de voordelen van geïntegreerd zoeken en tonen zelfs zoekmachines - Google voorop - in één resultatenlijst meteen wat ze in hun verschillende gespecialiseerde silo's gevonden hebben.

Maar tegenwoordig lijkt iedereen dol op gespecialiseerde Apps. Prima hoor. Ze zien er vaak ook prachtig uit, zoals Flipboard waar die Informatie Curator in zat. En als ik wil weten of het gaat regenen, maakt het mij ook niet uit of ik dat in een Buienradar-App kan zien. In mijn browser moet ik tenslotte ook buienradar.nl opvragen. En als ik die App niet heb, dan kan dat ook nog gewoon in een browser. En ook gewoon op mijn PC, als ik geen apparaat heb waar die App op werkt. Maar zulk multiplatform gebruik is lang niet voor elke toepassing mogelijk. En zelfs op de PC moet je voor bepaalde functies steeds vaker een speciale browserextensie of add-on installeren.

Zo word je haast gedwongen toch weer een tweestaps proces te doorlopen als je informatie zoekt. Eerst een zoekactie doen welke Apps in aanmerking komen om je informatiebehoefte te bevredigen. En dan in elk van die Apps nog eens kijken of er wat uitkomt. Gelukkig biedt Google voor die eerste stap wel enig soelaas in de vorm van facet-zoeken. Afgekeken van de interfaces van onze Discovery tools? Niet wat betreft de presentatie, want je moet bij Google wel erg zoeken naar die optie. Maar als je weet waar die verborgen zit, kun je een totaalzoekresultaat wel met één klik inperken op alleen Appjes. Maar echt facetzoeken blijkt dat toch niet, want ik kreeg die keuze ook in de situatie waarin Google als resultaat van mijn inperking meldde "Your search - ***curation tools*** - did not match any documents". Weg dus met die Apps. Gewoon weer alles in mijn browser (op welk apparaat dan ook).

Weg met Google

Af en toe bekruipt me het gevoel dat Google maar eens moet ophoepelen, want ik word helemaal gek van ze. Een van hun laatste streken is dat de "translated foreign pages search" plotseling verdwenen was. Bij het geven van de VOGIN-cursus kwamen we daar achter. Een paar dagen eerder was die optie er nog. En het heeft toen nog dagen geduurd voor officieel bekend werd dat het definitief was opgedoekt en nooit meer zou terugkomen.

Het was een geweldige functie. Je tikte je zoekvraag in je eigen taal in. Je selecteerde de taal waarin je wilde dat gezocht werd. Jouw zoekwoorden werden in die taal vertaald en daarmee werd de zoekvraag uitgevoerd. De lijst resultaten (in die taal) werd vervolgens automatisch naar je eigen taal terugvertaald. En als je een resultaat aanklikte, werd de inhoud van die pagina ook naar je eigen taal vertaald. Uiteraard niet allemaal foutloos, maar een prachtig hulpmiddel om een indruk te krijgen van hetgeen over een onderwerp werd geschreven in talen die je zelf niet kent. Dat ging heel wat verder dan een toevallig met eigen zoekwoorden gevonden pagina in het Swahili door Chrome laten vertalen, wat nu nog wel kan.

Maar het bleek onze eigen schuld. We gebruikten het volgens Google niet voldoende. Geen wonder denk ik dan, als Google die functie steeds dieper in zijn menustructuur wegstopt, zodat je hem niet meer vindt.

Het meest braakwekkend was nog Google's eigen commentaar. Ze hadden niet een schitterende functie om zeep gebracht, nee het ging alleen om *"streamlining"* om nog beter in staat te zijn *"to focus on creating beautiful technology that will improve people's lives."*

Terwijl dit nog maar net bezonken was, ontdekte een collega al weer iets anders. Bij Google Scholar kun je nog maar maximaal 20 resultaten tegelijk in je resultatenlijst te zien krijgen, terwijl echte informatieprofessionals dat toch meestal op 100 hadden staan. Maar ja, Google weet wat goed voor ons is. Het zou me niet verbazen als Google Scholar zelf op een kwade dag ook ineens verdwenen blijkt. Het is al enige tijd uit alle uitklapmenus op Google's homepage verdwenen. Gering gebruik is zo al voorgeprogrammeerd.

Terwijl iedereen half mei aan de sociale media hing om maar niets te missen van alle "beautiful new technology" die Google-techneuten op hun I/O event aankondigden, verdween dus tegelijkertijd onaangekondigd nuttige functionaliteit. Eerder hadden we (wel aangekondigd) natuurlijk al het einde van Google Reader gehad. Maar dat lag toch anders: daarvoor bestaan allerlei alternatieven van andere aanbieders. Die verdwenen zoekfunctionaliteit was uniek voor Google; daarvoor kun je niet naar de concurrent.

Misschien moeten we Google maar eens massaal boycotten, want voor het zoeken dat ze nog wel bieden, bestaan genoeg andere zoekmachines. Niet dat ze daardoor op andere gedachten komen, want een paar dissidente informatieprofessionals zullen in Google's gebruikscijfers heus niet terug te vinden zijn. Maar het voelt gewoon lekker.

De mythe van de markt

Ik heb hier al vaker over Omega geschreven, de in Utrecht ontwikkelde voorloper van de huidige Discovery Tools. De systemen waarmee nu bijna alle universiteitsbibliotheken en steeds meer hogescholen alle digitale informatie waarvoor ze licenties hebben, via één interface (en één index) doorzoekbaar maken. Twaalf jaar geleden moesten we dat in Utrecht nog zelf ontwikkelen. Nu zijn er gelukkig marktpartijen die dit kant-en-klaar aanbieden. De voordelen zijn duidelijk. Grote spelers met veel klanten kunnen ontwikkelkosten over veel meer organisaties spreiden en zijn beter in staat om met alle relevante contentleveranciers te regelen dat ze hun metadata kunnen indexeren. En zelfs de full-text kan zo doorzocht.

Maar dan komt de markt echt in actie. Sommige leveranciers van Discovery Services produceren zelf ook metadata. En dan wordt concurrentie ineens erg contraproductief, want andere Discovery Services krijgen die metadata dan lekker niet. Zo vervliegt de droom van single search in alles waartoe je toegang hebt. Wat in Utrecht met Omega net niet lukte door gebrek aan middelen, blijkt de markt niet te lukken uit concurrentieoverwegingen. Een NMA-tje misschien?

Amerikaanse bibliotheken die klant waren bij zowel Ebsco als Primo, hebben vorig jaar nog even gedreigd hun abonnementen alleen te continueren als die twee tot een schikking zouden komen, zodat metadata van de één wel gebruikt mochten worden in het zoeksysteem van de ander. Maar als klant sta je hierbij per definitie zwak. Commerciële informatieleveranciers hebben nu eenmaal een monopoliepositie. Je kunt ze niet boycotten als toegang tot hun informatie voor jouw organisatie essentieel is.

Deze zomer vlamde die discussie op Twitter ook weer op. OCLC heeft een contract met Kluwer afgesloten om juridische bronnen in Picarta op te nemen. Daarmee kan een Kluwer-abonnee die ook ter beschikking krijgen in Worldcat Local, als ie dat als Discovery Tool gebruikt. Maar ja, daar heb je niets aan als jouw organisatie toevallig Primo Central, Summon of Ebsco EDS heeft gekozen. En eisen stellen helpt niet, zagen we al.

Hoe gaat het intussen met het oude Utrechtse Omega? Daar wordt 1 september de stekker uitgetrokken. "Omega unplugged" zoals @UBAbert op Twitter grapte. Last van de wet van de remmende voorsprong en omdat verder ontwikkelen te duur wordt? Inderdaad! Omdat nu een Discovery Tool van een marktpartij gekozen is? NEE! Afgezien van het feit dat nog altijd geen van die tools volledigheid kan garanderen, blijken al aanwezige systemen prima alternatieven, met een heel behoorlijke dekking. Voor wetenschappelijke artikelen zijn dat Google Scholar en de database Scopus. Google Scholar gratis, maar wel nog wat problemen met link-resolving om bij de juiste fulltext versie te komen (en nog wat andere Google-bezwaren). Scopus met prima linking, maar wel iets minder volledig (en Elsevier-duur). Heeft de markt dus toch gewonnen? Misschien wel een beetje, maar wel op een andere manier dan ik ooit verwacht (en gehoopt) had.

Het einde van de telegraaf

Weet u nog wanneer u voor het laatst een telegram verstuurd hebt? Vast niet, want dat moet al heel lang geleden zijn geweest. Dat realiseerde ik me toen ik las dat op zondag 14 juli het allerlaatste telegram verstuurd zou zijn. Dat was in India. Die laatste dag hebben nog een heleboel Indiërs telegrammen naar hun minister voor Telecommunicatie gestuurd, met het verzoek de dienst te continueren. Maar dat heeft geen effect meer gehad. De Indiase overheid schijnt zo ongeveer de laatste instantie te zijn geweest die nog regelmatig telegrammen gebruikte om met zijn onderdanen te communiceren. In de VS en ook in Nederland bestond de dienst al lang niet meer. Toch was de eerste T in "PTT" nog steeds die van Telegrafie gebleven, zelfs toen "de" PTT allang was opgesplitst in PTT-post en KPN (Koninklijke PTT Nederland).

Zelf herinner ik me het gebruik eigenlijk nog alleen van gelukstelegrammen bij huwelijk of geboorte. In het kader van huidige discussies vroeg ik me nog af hoe het eigenlijk zat met de privacy-aspecten van die dienst. In elk geval bestond in de praktijk niet zoiets als telegram-geheim. Te verzenden tekst moest je namelijk inleveren bij een telegraafkantoor - meestal een speciaal loket op een postkantoor - om hem daar door een beambte te laten overtikken.

Op het postkantoor in de plaats van bestemming kwam de boodschap dan weer te voorschijn, vaak nog uitgeprint op een smal papieren lintje. Dat moest dan weer door iemand in regels worden opgeknipt en op een vel papier geplakt om het zo bij de geadresseerde thuis te laten bezorgen. Niet echt een medium voor het plannen van bankroof of terroristische aanslag. Dat maakte dat er toen (voorzover ik weet) nog geen big brother nodig was om alle telegraafverkeer wereldwijd te scannen op mogelijk terroristische neigingen.

Hoeveel telegrammen zouden ooit verstuurd zijn? Ook dat zal nooit centraal zijn bijgehouden. Dat moeten er heel wat minder zijn geweest dan de 540 miljard tweets die intussen wel centraal in de Topsy zoekmachine zitten en in één keer allemaal doorzoekbaar zijn. Zeker als ik afga op mijn eigen communicatiegedrag - 4 telegrammen versus 4000 tweets. Toch werden in de toptijd in toplanden als de VS en India jaarlijks 200 miljoen telegrammen verstuurd. Zelfs in 160 jaar haal je daarmee die 540 miljard nog lang niet, al scheelt het niet zoveel ordes van grootte als ik had verwacht.

Wanneer zouden we een soortgelijk bericht kunnen verwachten dat de allerlaatste e-mail is verstuurd? Nu kunnen we ons nog moeilijk voorstellen dat e-mail achterhaalde technologie wordt, maar dat konden telegram-gebruikers 60 jaar geleden ook nog niet. Toch is de telegraaf nu wel definitief helemaal uit de tijd. Die met een kleine letter t bedoel ik. Hoewel, ... misschien geldt dat voor die met hoofdletter T langzamerhand ook wel.

Big, bigger, biggest

In een presentatie "How search works - from algorthms to answers" meldde Google vorig jaar intussen 30 biljoen (in het Engels 30 trillion) verschillende URL's geregistreerd te hebben. Zulke getallen gaan het menselijk voorstellingsvermogen ver te boven. Daarom was ik maar eens gaan rekenen. Als je elk van die URL's op een apart velletje papier zou opschrijven en die netjes zou opstapelen, zou dat acht stapels papier opleveren, elk van hier tot aan de maan. Toen ik kortgeleden die presentatie nog eens bekeek, zag ik dat Google dat aantal URL's intussen tot 60 biljoen had verdubbeld. Dat rekent makkelijk: nu dus 16 stapels tot aan de maan. Dat is nog eens *big data*.

Maar wees gerust: gelukkig zijn die webpagina's niet allemaal vindbaar. Hoewel Google daarover geen cijfers bekend maakt, zijn er schattingen dat hun index enkele honderden miljarden webpagina's doorzoekbaar maakt. Dat is nog geen procent van al die URL's. Kennelijk is men goed in staat enorme aantallen identieke pagina's, spam en linkfarms als zodanig te herkennen en uit te sluiten. Toch is ook dat aantal pagina's nog aardig *big*.

Big data zijn intussen hot. Alle grote bedrijven schijnen ze te hebben en er iets mee te willen. Onlangs werd in een kritisch Gartner-rapport onomwonden gesteld dat bedrijven die het meest de mond vol hebben over hun *big data* projecten, vaak het minst idee hebben wat ze er eigenlijk mee willen. En inderdaad, zolang op veel plekken zelfs *small data* nog nauwelijks terug te vinden is, vraag je je af wat ze dan met *big data* moeten.

Intussen werken we natuurlijk allemaal hard mee om *big data* te genereren. Door ons zoek-, klik-, browse-, tweet- en like-gedrag dat overal geregistreerd wordt. Daarnaast door alle foto's en video's die we maken, doorsturen, uitwisselen, opslaan in de cloud, en voor alle zekerheid nog een paar keer copiëren. Dat maakt dat al een hele tijd onze dataproductie elk jaar bijna verdubbelt. En dat gaat nog wel even door als ook onze huishoudelijke apparaten in het "internet of things" hun gegevens gaan toevoegen. Zo'n jaarlijkse verdubbeling betekent dat we elk jaar evenveel bytes produceren als alle voorgaande jaren samen. Toch moet dat eens ophouden. Anders produceren we in het jaar 2113 evenveel bytes als er atomen in onze aardbol zitten.

Er zijn dus natuurlijke "grenzen aan de groei". Ruim voor 2113 zullen we al moeten ophouden met die jaarlijkse verdubbeling van onze productie. En we moeten ook niet meer alles wat we produceren willen bewaren. Archivarissen zijn van oudsher goed in het vaststellen van selectiecriteria wat wel en niet bewaard moet worden. Dat moeten ze ons gewone mensen ook maar eens leren, zodat we niet meer allemaal onze volledige digitale voetafdruk tot in de eeuwigheid duurzaam willen opslaan.

Queries als kunst

Er is een "Society of the Query". Dat is geen genootschap, zoals de naam suggereert, maar een internationaal congres dat begin november voor de tweede keer in Amsterdam gehouden werd. Zoekmachines staan daarin centraal, maar op een heel andere manier dan wij als informatie-professionals gewend zijn. Zo was er ook nu weer een track over zoekmachines en Kunst (ja met grote K), waarin het onder meer ging over een project waarin Google's *similar images* een rol speelden. Maar verder ging het vooral over sociale, politieke en filosofische aspecten van onze zoekmaatschappij.

Een Belgische filosofe (Antoinette Rouvroy) hield de toehoorders voor dat met de aandacht voor *big data* het idee van causaliteit verloren is gegaan. Men denkt dat kennis al in die data verborgen zit en *"only has to be surfaced by magic"*. Ook was er een bijdrage (Kylie Jarrett) over de *Metaphysics of search*. Een andere spreker Anton Tantner) besprak zoeken in historisch perspectief. Het Orakel van Delphi, het Domesday Book, Parijse concierges en de vooroorlogse romanfiguur butler Jeeves plaatste hij op één lijn met Google. Elk was in zijn eigen periode ***de*** bron waaruit men essentiële informatie putte.

Kritische opmerkingen over Google of Facebook werden hier uiteraard niet uit de weg gegaan. Zeker niet door keynote speaker Siva Vaidhyanathan, auteur van het boek *"The Googlization of Everything (and Why We Should Worry)"*.

Google is voor hem - net als voor ons - een blackbox waarvan de precieze werking onbekend is, maar wel voortdurend verandert. Tegelijk is Google het meest succesvolle surveillance-systeem op aarde geworden. Je mag je dus afvragen waarom wij als gebruikers geneigd zijn onze data eerder aan Google toe te vertrouwen dan aan de NSA.

Google-baas Eric Schmidt heeft in 2010 al eens gezegd dat Google beter onze eigen gedachten kende dan wijzelf. Volgens Vaidhyanathan zijn Google, Facebook en Apple er dan ook niet langer op uit het besturingssysteem voor onze computers te leveren, maar willen ze het besturingssysteem van ons hele leven zijn. Superieure zoekalgoritmen staan voor Google niet meer voorop; het gaat ze om onze gegevens. Bij dat eerste mag je misschien een vraagteken zetten, want dat superieure zoeken lijkt toch nog steeds nodig om ons te verleiden - al is het indirect - die data van ons leven beschikbaar te stellen. Maar volgens Siva wil Google wel al opties achter de hand hebben als alternatief voor hun zoekmonopolie. Daarom ontwikkelen ze Google Glass en hun zelf-rijdende auto.

In dat perspectief dringt zich de vraag op of wij als informatieprofessionals ook in de toekomst nog op Google (of andere zoekmachines) kunnen bliijven rekenen als hulpmiddel om onze informatie op te sporen. Misschien moet Siva Vaidhyanathan zijn verhaal ook maar eens voor ons publiek houden. Voorlopig lijkt Google's blackbox het zoeken toch nog wel even te blijven domineren. Zolang zal het opzetten van goede queries ook nog een hele kunst blijven.

* Society of the Query #2 – Conference Report (2013) - http://networkcultures.org/query/past-events/2-amsterdam/

Op mijn zeepkist in de woestijn

"Ik voel me net of ik op een zeepkist in het park sta te roepen" sprak mijn vriendin toen ze zich destijds door mij had laten verleiden ook op Twitter actief te worden. In Hyde Park zie je tenminste nog of iemand blijft staan luisteren. Op Twitter heb je geen idee of je berichten ook maar door iemand worden opgemerkt, zeker als je nog bijna geen volgers hebt. Hoewel altijd tot tegenspreken geneigd, moest ik haar hierin wel gelijk geven. Ook al zat ik zelf in een wat luxere positie, doordat ik via mijn toegang tot de oude media al snel flink wat volgers had weten te verzamelen.

Toch begint het bij mij ook wel te knagen. Ik merk steeds vaker dat een van mijn volgers een berichtje plaatst of retweet, dat ik zelf al langer dan een week geleden heb gemeld. Maar ik blijk niet degene die geretweet wordt. Mijn bericht hebben ze kennelijk helemaal niet gezien. Sta ik, ondanks die 900 volgers, toch nog altijd een beetje in de woestijn te roepen? Eigenlijk is dat heel voorstelbaar. Zelf volg ik - heel selectief - maar zo'n 70 twitteraars. En toch heb ik vaak al moeite om mijn tijdlijn bij te houden. Hoe moet dat dan zijn als je er 700 of misschien wel 1700 volgt?

De problematiek van het gelezen worden leeft kennelijk wel. Terwijl ik al aan deze column was begonnen, kwamen daarover nog diverse berichten langs. Het AllthingsD e-Zine berichtte in vette letters **"*(Almost) No One Is Reading Your Tweets*"**. Daarbij baseerden ze zich op een rapport van O'reilly dat vooral benadrukte dat er zoveel mensen zijn die nauwelijks volgers hebben. De mediaan schijnt op één volger te staan! Ja, dan word je zeker niet gelezen. Nog recenter blogde (en twitterde) Jeanine Deckers onder haar schuilnaam @Tenaanval over het *"opblazen van je tijdlijn"*, waarbij dat woord "opblazen" in destructieve zin bedoeld was. Zij beriep zich op ene Charlie Warzel die in Buzzfeed opriep **"*I Nuked My Twitter Feed And You Should Too*"**. Het ergerde hem dat zijn feed vaak sneller over het scherm raasde dan hij überhaupt kon lezen. Het aantal door hem gevolgden had hij daarop resoluut van 1838 naar 1 teruggebracht.

Mijn volgers volgen dus kennelijk ook te veel. En dan knaagt het aan mijn pretenties als **MIJN** berichten niet worden gelezen. Maar in feite doet dat er natuurlijk niet toe. Zolang de echt interessante nieuwtjes en opinies maar blijven rondzingen. Als we die maar blijven herhalen, dan krijgt iedereen in ons kringetje ze uiteindelijk wel een keer onder ogen, ook wie zijn tijdlijn nog niet heeft opgeblazen, ook wie toch nog eigenwijs probeert om 1700 tweeps te volgen. Laten we dus nog maar even blijven roepen op onze zeepkisten in de woestijn. Dan blijft Twitter de uiterst nuttige attenderingsdienst die het nu is.

* Tweets loud and quiet; Twitter's long, long, long tail suggests the service is less democratic than it seems (O'Reilly 2013) - https://www.oreilly.com/ideas/tweets-loud-and-quiet

ZMO voor optimale semantiek

ZMO heeft in Nederland nooit ingang gevonden als afkorting voor ZoekMachineOptimalisatie. Iedereen heeft het hier altijd over SEO; ik ook. Eigenlijk is dat een raar begrip - zowel in het Engels als in het Nederlands. Zoekmachines zelf worden daarmee namelijk helemaal niet geoptimaliseerd. Contentleveranciers - als we in dit verband de makers van webpagina's even zo mogen noemen - hebben helemaal geen directe toegang tot die zoekmachines. Als ze die zoekmachines toch een beetje willen *tweaken*, zijn ze dus wel gedwongen om indirecte maatregelen te nemen, via de inhoud en de links van hun eigen pagina's.

Als gebruiker van zoekmachines vraag ik me wel eens af wat ik eigenlijk aan dat optimaliseren heb. Zorgt het voor misleiding van mij als zoeker? Vind ik daardoor dingen die een ander wil dat ik vind, in plaats van wat ik zelf wil? Of zorgt het juist voor betere vindbaarheid? Als het goed is toch vooral het laatste. En dat geldt ook in het huidige tijdperk van semantisch zoeken. Toepassing van semantische technieken wordt zelfs meer in SEO-context aan de man gebracht, dan in verband met zoekvaardigheden. De commercie wordt dringend aanbevolen semantische coderingen - dus metadata - in de code van hun webpagina's te verwerken, opdat die door zoekmachines beter gevonden en hoger gerankt worden. En om te zorgen dat ze daardoor met extra "rich snippets" in resultatenpagina's verwerkt worden, op basis waarvan zoekers beter en sneller kunnen beoordelen wat de resultaten zijn waar ze eigenlijk naar op zoek zijn.

Intussen rukt SEO ook op in de wetenschappelijke wereld. Tijdschriftuitgevers geven wetenschappers tips hoe ze hun artikelen zo moeten schrijven dat ook die beter gevonden worden. Tips voor het gebruik van titels waarin de inhoud ondubbelzinnig tot uiting komt. Dus geen overdrachtelijke formuleringen, woordspelingen en grappen meer. Saai is beter. En ook de tip om in de samenvatting van je artikel volop te variëren met allerlei synonieme termen voor de onderwerpen waarover je het hebt. Sommige universiteiten organiseren voor hun promovendi zelfs cursussen daarvoor. Van artikelen die beter gevonden kunnen worden, wordt impliciet verwacht dat ze ook vaker worden gelezen en dan hopelijk ook vaker geciteerd.

Als auteurs zich aan die richtlijnen houden, zou het voor ons zoekers dus makkelijker moeten worden. Of we nu het ene woord als zoekterm gebruiken of een synoniem ervan, (alle?) relevante artikelen vinden we toch wel. Dat betekent eigenlijk dat auteurs wordt geleerd zo te schrijven dat hun artikelen niet meer met gecontroleerd vocabulaire ontsloten hoeven te worden. Zodat we ook in Google Scholar (of zelfs in de gewone Google) hun artikelen makkelijk kunnen vinden. Adieu thesauri, adieu classificaties en taxonomieën, adieu gestructureerde bibliografische databases. Auteurs worden verantwoordelijk voor hun eigen ZMO en moeten zelf voor hun semantiek gaan zorgen. Met dezelfde drijfveer als bij de commercie.

Infobesitas in de 19de eeuw

Wij denken altijd dat wij het nu zo moeilijk hebben met onze informatieovervloed. Maar ik kwam er onlangs achter dat men dat 120 jaar geleden ook al zo voelde. In mijn voorjaarsvakantie had ik me verdiept in een studie van Auke van der Woud over het ontstaan van het moderne Nederland ("Een Nieuwe Wereld"). Dat boek gaat vooral over de grote technische en maatschappelijke veranderingen in de tweede helft van de 19de eeuw. Daarin schrijft hij onder meer (en ik citeer nu):

"De aandacht die voor het geduldig oefenen en ontwikkelen van het denken en aanvoelen nodig was, werd afgeleid door een steeds sterkere stroom van actuele nieuwsfeiten die men niet mocht missen, op straffe van het achterop lopen, van het niet *bijblijven*."

Het lijkt wel of ik Nicholas Carr hoor, die in diens boek "The Shallows" soortgelijke klachten ventileert over onze huidige digitale maatschappij.

Van der Woud haalt ook de Duitse medicus en zionist Max Nordau aan. In diens boek "Entartung" uit 1892 - dat vooral over heel andere dingen ging, zoals de omineuze titel al suggereert - schreef hij dat door de enorme toename van de communicatie en mobiliteit, de wereld van een gemiddelde dorpsbewoner al veel groter was dan die van een eerste minister honderd jaar eerder.

> *"Die dorpsbewoner praat mee over de hongersnood in Rusland, een opstand in Chili, het Panamaschandaal en een wereldtentoonstelling in Amerika. De correspondentie van een keukenmeid is tegenwoordig groter dan die van een professor een eeuw geleden."*

En dan denken wij dat dat allemaal pas door internet is gekomen. Of zou Nordau toch een iets te optimistisch beeld gehad hebben van die dorpsbewoners en keukenmeiden uit 1892? Net zo goed als wij van die uit 2013. Of bestaan er intussen helemaal geen dorpsbewoners en keukenmeiden meer?

Die groei van de communicatie waar Van der Woud en Nordau het over hebben, was vooral veroorzaakt door de opkomst van het telegraafverkeer. Daarvoor lagen in 1895 al veertien transatlantische kabels op de bodem van de oceaan tussen Europa en Amerika. De voorlopers van de veelvoud aan veel snellere kabels die nu alle continenten verbinden. En dan te bedenken dat net dit jaar het telegraafverkeer officieel wereldwijd is beëindigd.

Nog een aardig punt: Van der Woud bespreekt ook 19de eeuwse klachten, dat men voortdurend van het echte werk werd afgeleid, omdat wel vijf of zes keer per dag een postbesteller brieven kwam bezorgen, telkens als weer een volgende trein met post op het station was aangekomen. Zulke klachten klinken ook weer heel vertrouwd. Maar ook wel navrant dat onze postbestelling dezer dagen juist tot 4x per week dreigt te worden ingeperkt. Over een eeuw dus nog maar 4x per week bezorging van onze email - of zelfs helemaal afgeschaft, net als de telegraaf? Maar ook dan zal de infobesitas nog niet over zijn, durf ik wedden.

Metadata is niet meer wat het geweest is

Metadata zou best eens het woord van 2014 kunnen worden. Iedereen heeft het er ineens over. Vijf jaar geleden placht ik mijn IDM-studenten nog omstandig uit te leggen wat metadata waren. Alleen voorlijke automatiseerders gebruikten het woord toen al een tijdje. En een paar pretentieuze bibliothecarissen die *metadata* chiquer vonden klinken dan het oubollige *titelbeschrijving*. Daartoe geïnspireerd door Amerikaanse vakgenoten die in personeelsadvertenties al langer *Metadata Librarians* en *Metadata Coordinators* wierven.

Dat het woord 'metadata', behalve bij wat vakidioten, hier vrijwel onbekend was, blijkt ook wel uit het voorkomen ervan in de Nederlandse Krantenbank. In de periode tot en met 2013 kwam het maar tien keer voor in koppen van artikelen, waarvan zelfs maar drie keer in gewone kranten. De rest kwam uit vakbladen, zoals onze eigen IP, de VIP en InCT. Maar alleen al in de eerste drie maanden van dit jaar komen we het 30 keer tegen. Dankzij Edward Snowden en de afluisterpraktijken van de NSA heeft ineens iedereen het erover en iedereen denkt te weten wat metadata zijn. Zo meldde de Volkskrant onder meer: "Ook Plasterk wist al maanden over metadata"; ja, hij wel. En ook nog: "NSA verzamelt 6 miljard metadata per dag".

Dat is nog wel van een andere orde dan de metadata waarover begin dit jaar op OpenGLAM bericht werd: "Nearly all German National Library metadata now available under CC0 license". Het heeft brave bibliothecarissen waarschijnlijk meer dan een eeuw gekost om die tien miljoen metadata te verzamelen. Maar daartoe heeft dan wel iedereen vrije toegang. En dan nog een opmerkelijk verschil. Intussen hebben wij voor onze zoeksystemen meestal een voorkeur voor full-text zoeken, boven alleen metadata. Maar de NSA beweert voor hun analyses wel met alleen maar metadata toe te kunnen. Of zouden ze stiekem toch alles bewaren wat wij zeggen en schrijven om daarop datamining los te laten (zoals Google natuurlijk al lang doet)?

Uit historisch besef heb ik ook nog even gekeken in de gedigitaliseerde oude kranten van de KB (die nu in de nieuwe dienst Delpher zitten). Vijftien vindplaatsen voor 'metadata'! Maar veertien daarvan, uit de jaren 1894 - 1934, blijken voort te komen uit de meest onwaarschijnlijke OCR-fouten. In werkelijkheid bleek daar *matador, metselaar* of *margarinefabriek* te staan. Alleen in een overzicht van radio- en tv-programma's in het Limburgs Dagblad kwam echt het woord metadata voor. Op zondag 2 februari 1975 zond de Belgische TV een tien minuten durend filmpje uit dat kortweg "Metadata" heette. Dat maakt wel heel nieuwsgierig wat onze Belgische buren daar bijna 40 jaar geleden al over te melden hadden.

Nog even iets recenters. In mijn bookmarks zat ook nog een bijna twee jaar oude blogpost met de titel: "Metadata isn't what it used to be". Dat ging toen nog over catalogiseren, maar buiten die context klinkt deze verzuchting nu wel heel actueel.

Verzamelingenleer

Verzamelen van *dingen* schijnt een algemeen menselijk behoefte te zijn. En zelfs bij dieren kom je het verschijnsel tegen. Pack-rats zijn daar een mooi voorbeeld van. Af en toe worden door archeologen nog "verzamelingen" ontdekt die ooit in het Pleistoceen door deze diertjes hun holen zijn binnengesleept.

Ik kom hierop omdat ik laatst ergens iets *leuks* moest vertellen over verzamelen. Als zoekspecialist vraag ik me dan af of het problemen oplevert als ik bij Google naar zo'n onderwerp ga zoeken. De zoekterm "verzamelen" geeft nauwelijks moeilijkheden - behalve dat je 6 miljoen resultaten krijgt. Maar met "verzameling" ligt dat anders. Als echte bèta denk ik daarbij ook meteen aan verzamelingenleer, een nogal axiomatisch onderbouwd specialisme uit de wiskunde. Toch wordt ook daar een verzameling simpelweg gedefinieerd als "een collectie van verschillende objecten, elementen genoemd, die op haar beurt ook weer als een object wordt beschouwd". Maar of je nu erg gelukkig wordt als het zoeken naar verzamelingen allerlei wiskundige verhandelingen oplevert over Venn-diagrammen, doorsneden, kardinaalgetallen, machtsverzamelingen en de *lege verzameling*?

In het Engels blijkt het nauwelijks makkelijker te gaan. Met "collect" of "collecting" gaat het nog wel goed, maar met "collections" krijg je problemen. Niet met verzamelingenleer, want die verzamelingen heten in het Engels "sets", maar wel met het feit dat ook elke bibliotheekcollectie een "collection" heet.

Inderdaad ook verzamelingen, al noemen wij die in het Nederlands nu juist weer niet zo. "Collection management" zal ons dus zelden leren hoe je gewone verzamelingen moet beheren, maar vooral hoe je met bibliotheekbezit moet omgaan.

Zelf heb ik ook een heleboel verzamelingen. Maar misschien is verzamelen daarvoor niet het juiste woord en gaat het bij mij meer om "niet weggooien", ongeveer net zoals bij die pack-rats. En dan heb ik het nog niet eens over digitale verzamelingen. Daarvan gaat de groei meestal vele malen sneller dan bij fysieke collecties. Bijvoorbeeld omdat je zo makkelijk digitale foto's maakt. En digitaal hebben we nog minder neiging ooit iets weg te gooien. Vorig jaar (november 2013) schreef ik al eens dat, als de huidige groei van onze dataproductie aanhoudt, we over een eeuw evenveel bytes produceren als het aantal atomen waaruit onze aarde bestaat. De Club van Rome zei het 40 jaar geleden al: er zijn "grenzen aan de groei". We moeten misschien maar eens gaan ontzamelen. De "Inside Higher Ed" blog zei deze week zelfs over bibliotheken al dat "*Most of us would improve our collections enormously if we got rid of lots of books*".

In mijn eerdere wiskundige uitstapje noemde ik even de *lege verzameling*. Dat is wel het ultieme resultaat van *ontzamelen*. En het is in elk geval een verzameling die we allemaal compleet hebben. Dat geeft voldoening. Behalve dat de wiskunde roet in het eten gooit. Volgens de verzamelingenleer kan er maar ÉÉN lege verzameling bestaan. En wie van ons mag die dan hebben?

* "Allemaal verzamelen"; een video-column - https://www.youtube.com/watch?v=wX5uWPMNtBM

Tegen het vergeten

Er is al heel wat afgeschreven over de "Right-to-be-Forgotten"-uitspraak van het Europese hof en wat dat voor Google betekent. Vooral Amerikanen vinden het volstrekt onbegrijpelijk en beschouwen het als censuur. Zij hebben de vrijheid van meningsuiting immers hoog in hun vaandel staan - wel gecontroleerd door de NSA natuurlijk. Maar het voelt ook alsof een bibliotheek alle boeken uit zijn catalogus zou moeten verwijderen, waarin achterhaalde informatie staat. Maar ze mogen op de plank blijven staan. Op basis van dat criterium zouden we nog wel even werk hebben. En ook zonder beroep op de vrijheid van meningsuiting moet voor rechtgeaarde informatieprofessionals natuurlijk alle informatie vindbaar blijven.

Overigens benadruk ik nog maar eens dat er niet echt informatie verdwijnt. De betwiste pagina's zelf blijven gewoon bestaan en Google haalt ze ook niet uit zijn index. Op google.com blijven ze zelfs gewoon in zoekresultaten verschijnen. Alleen op Europese versies als google.fr, google.de of google.nl worden betwiste pagina's uit de resultatenlijst verwijderd. En dat meldt Google ook onderaan: "*Some results may have been removed under data protection law in Europe*". Maar die melding wordt niet erg selectief toegepast. Hij blijkt te verschijnen bij elke zoekvraag die door Google als persoonsnaam wordt geïnterpreteerd. Ook wie bij google.nl naar "Eric Sieverts" zoekt, krijgt deze melding, terwijl ik echt niets heb proberen te verdoezelen van wat op het web over mij te vinden is.

Intussen worden al regelmatig onbedoelde voorbeelden gemeld: verwijderde berichten over bedriegers en zwendelaars dat ze bedriegers en zwendelaars waren - en waarschijnlijk nog zijn, zodat we toch wel graag gewaarschuwd willen blijven. En sowieso gaat Google hier klungelig mee om. Misschien ook wel expres, opdat maar goed duidelijk wordt hoe onuitvoerbaar die RtbF-wet is.

Bovendien lijkt die ook nog het tegendeel te bevorderen van wat beoogd werd. Verwijderde informatie wordt soms juist beter vindbaar. Diverse weblogs wezen er al op dat ook hier het "Streisand-effect" optreedt. Genoemd naar Barbra, die - uit angst voor inbrekers - een foto van haar dure villa van internet verwijderd wilde hebben, waarna die foto duizendvoudig op allerlei sites verscheen. Zo weet nu ook de hele wereld dat die Spaanse meneer ooit failliet ging. En media als Guardian en BBC rapporteren uitgebreid over oude berichten waarvan Google ze meedeelde dat ze waren verwijderd. Bovendien is iemand een website "hiddenfromgoogle.com" begonnen, waarop wordt bijgehouden wat er allemaal verwijderd wordt. Daar iets vinden is weliswaar niet zo makkelijk als bij Google, maar dit soort meldingen en berichten maakt automatisch dat in Google die oude informatie juist weer vaker komt bovendrijven, ondanks weglaten van de oorspronkelijke pagina.

Of moeten we het probleem misschien toch wat relativeren? Door de overmaat aan informatie op internet dreigen relevante resultaten over vrijwel elk willekeurig onderwerp toch al steeds meer onder te sneeuwen. Misschien dus maar goed dat Google dingen vergeet? Nee! Vergeet het maar. Geen geforceerde Alzheimer.

Ontsluiting terug van weggeweest

Ontsluitingssystemen als thesauri en classificaties zijn een beetje uit de mode geweest. Net als gestructureerde databases. Met Google kun je immers alles vinden, ook als informatie niet gestructureerd en niet ontsloten is. In dat perspectief is het opmerkelijk dat juist Google de laatste jaren het roer heeft omgegooid. Ze kunnen natuurlijk nog altijd goed in ongestructureerde informatie zoeken, maar daarnaast leggen ze meer nadruk op semantiek. Semantiek door structuur en door wat wij ontsluiting plegen te noemen. Liefst verwerkt in "gestructureerde markup" in HTML-code.

In ongestructureerde informatie worden trouwens ook zogenaamde "impliciete" entiteiten herkend. Zowel Google als Bing gebruiken daarvoor gestructureerde kennissystemen. Bij Google beperkt zich dat niet meer tot de Knowledge Graph, met gestructureerde data uit onder meer de Wikipedia (de DBpedia). Als aanvulling is er nu ook een "Knowledge Vault" met kennis die via slimme technieken - Probabilistic Knowledge Fusion - uit ongestructureerde bronnen wordt geëxtraheerd. Dat "probabilistic" wil zeggen dat van elk gegeven een waarschijnlijkheid wordt berekend dat het waar is, en "fusion" dat gegevens uit allerlei bronnen gecombineerd worden. Zo kan steeds meer betekenis in webpagina's worden achterhaald.

Terug naar de expliciete ontsluiting die in webpagina's wordt gebruikt. Een geaccepteerde standaard daarvoor is Schema.org, een soort universele thesaurus die vooral door de grote zoekmachines Google, Bing, Yahoo en Yandex ontwikkeld en ondersteund wordt.

Ik noem het hier wel een thesaurus, maar dat is het niet echt. Je markeert hiermee de betekenis van al in webpagina's voorkomende namen, woorden, getallen en dergelijke. Dat iets de naam van een acteur is, dat dit de prijs van de getoonde camera is, dat dit de tijdschriftaflevering is, waarin een artikel is verschenen. Ja, ook bibliografische gegevens kunnen sinds vorige maand met Schema.org gecodeerd worden, door een aanvulling met " Support for Bibliographic Relationships and Periodicals".

Dat ik Schema.org een soort thesaurus noemde is overigens niet zo gek, want de daarin gedefinieerde entiteiten zitten in hiërarchische boomstructuren. Maar het gaat nog een stapje verder, want je kunt ook relaties markeren tussen in een webpagina aanwezige entiteiten. En sinds kort kunnen die zelfs "rollen" toebedeeld krijgen. Daarmee begint Schema.org op een echte ontologie te lijken. Nu noemen semantisch web enthousiasten eigenlijk alles al een ontologie. Alles wat structuur aanbrengt, en alles wat betekenis en betekenisrelaties vastlegt; dus zeker Schema.org, maar ook onze klassieke thesauri en classificaties.

Wordt al die codering nu al veel in webpagina's toegepast? In 36% van de resultaten van Google zoekacties schijnen deze codes voor te komen. Overigens een misleidend percentage, want aanwezigheid van codes bevordert ook hogere ranking. Ook een goede reden die markup toe te voegen. En uiteraard niet handmatig, want gegevens komen toch al bijna altijd uit een gestructureerde database of CMS, zodat die codes automatisch gegenereerd kunnen worden. Thesauri, classificaties en ook gestructureerde databases zijn dus weer helemaal terug van weggeweest. Al zitten er nu andere modieuzere etiketjes op.

* Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion (2015) - https://noon99jaki.github.io/publication/2014.kdd.pdf

Yahoo cureert geen content meer

Content curation, is dat niet ook een soort ontsluiting? Toen ik vorige maand schreef dat ontsluiting weer helemaal terug was, had ik daarbij inderdaad ook *content curation* kunnen noemen. Ook dat is al weer enkele jaren een buzzword. Op Google levert het 350.000 hits op; een veelvoud van wat je met "subject indexing" vindt.

Des te opmerkelijker dat onderwerpsgidsen of directories zoals die van Yahoo - toch prototypes van content curation - juist steeds minder gebruikt worden. Zo weinig dat Yahoo eind van dit jaar de stekker eruit trekt. Terwijl de Yahoo-directory toch de allereerste onderwerpsingang tot het web was. Nog voor de eerste echte zoekmachines opkwamen, hadden Jerry Yang en David Filo in 1994 hun systematische lijstje bookmarks als "Yet Another Hierarchical Officious Oracle" op het web gezet. Overigens schreven zij zelf die benaming toe aan de Yahoos, nare primitieve wezens die voorkomen in *Gulliver's reizen*.

Ook later was Yahoo nog enige tijd populairder dan de echte zoekmachines, omdat je er makkelijker iets mee vond dan met Lycos, AltaVista of Excite. Op een internationaal zoekcongres in 1998 werd zelfs beweerd dat de Yahoo-directory het van de zoekmachines gewonnen had (D.L. Wiley; zie ook column mei 1998).

Hoe anders is het gelopen na het verschijnen van Google. Zoals die de andere echte zoekmachines uit de markt gedrukt heeft, zo is dat ook gebeurd met steeds meer gecureerde onderwerpsingangen. Argus Clearinghouse, Profusion, Looksmart, DirectSearch, BUBL, Pinakes, Intute, Goshme, DUTCHess - wie kent die namen nu nog? Allemaal al geruisloos verdwenen. Meest recent - een paar maanden geleden - volgde Complete Planet, een op onderwerp ingedeeld overzicht van bijna 100.000 onderwerpszoeksystemen. En per 31 december dus de Yahoo Directory zelf. Die mededeling verstopte Yahoo - nogal beschamend na 20 jaar trouwe dienst - in twee regeltjes in een bericht dat verder hoogdravend over "Continued Product Focus" ging.

De Open Directory bestaat intussen nog wel. Maar hoeveel wordt die nog gebruikt? En hoe goed wordt die nog bijgehouden door zijn 89.824 (!) editors? En de WWW Virtual Library? En de Nederlandse Startpagina's? Ze bestaan nog, maar de achterliggende gespecialiseerde onderwerpsgidsen lijken steeds minder goed up-to-date gehouden te worden. En bezoekcijfers van Startpagina.nl blijken in tien jaar tijd toch ook met een factor 10 te zijn ingezakt.

Dit alles wekt de indruk dat content curation niet veel toekomst meer heeft. Maar het hangt er natuurlijk maar vanaf wat je precies onder die term verstaat en om welke schaal het gaat. Gaat het om het hele internet of alleen om heel specifieke segmenten en gebruikersgroepen, of zelfs alleen om de content van je eigen website waar je Schema.org op toepast? Voor dat eerste lijkt inderdaad geen plaats meer, zoals het stopzetten van Yahoo's curatie-activiteiten laat zien. Maar voor kleinschaliger toepassingen lijken er nog volop mogelijkheden. Zij het dat zonder duidelijke doelgroep of use case ook dat laatste moeilijk wordt.

* Deborah Lynn Wiley / Beyond Information Retrieval: Ways to Provide Content in Context (1998) - http://www.cs.uml.edu/~haim/teaching/iws/tirsaa/sources/Beyond_IR_Content_in_Context.html

Een vak leren

"Om het niveau van bibliothecarissen op te voeren is begonnen met de eerste opleiding tot bibliothecaris op hbo-niveau". Met die zin begon 14 oktober j.l. een artikel in *De Ware Tijd - Dé krant die Suriname leest*. En dat artikel ging verder met de verzuchting dat de nood daar erg hoog was vanwege de vergrijzing, want "Het is precies dertig jaar geleden dat de laatste bibliothecaris, afgestudeerd in Nederland, naar Suriname terugkeerde".

Een opmerkelijk bericht nu in Nederland - met dertig maal zo veel inwoners als Suriname - nauwelijks nog IDM-ers afstuderen. Net nu het reguliere HBO onderwijsaanbod in Nederland eigenlijk alleen nog uit veel bredere opleidingen bestaat, waarin vaak nog maar een klein deel van de competenties van de "pure" informatieprofessional aan de orde kan komen.

Opmerkelijk ook, omdat de universitaire opleidingen, zowel in Nederland als in Vlaanderen al evenzeer in zwaar weer verkeren. Dertig tot veertig jaar geleden plachten al mijn vakgenoten die iets met wetenschappelijke informatie van doen hadden, nog een Opleiding tot Wetenschappelijk Bibliothecaris te volgen bij de Universiteit van Amsterdam. Jaarlijks zaten daar de collegezalen vol met meer dan zestig postdoctorale "studenten".

Die generatie is ook hier aan het vergrijzen en voor een deel al met pensioen. Voor de - overigens zeer enthousiaste en kundige - jonge generatie vakspecialisten bij Universiteitsbibliotheken en HBO-mediatheken bestaat niet meer zo'n opleiding. Naast de kennis van hun eigen (oorspronkelijke) vakgebied, moeten zij de extra kennis over het informatievak - want dat is wel degelijk ook een vak - op andere manieren bij elkaar sprokkelen.

Manieren daartoe zijn er natuurlijk wel, met modules en cursussen van GO-Opleidingen, met de VOGIN-cursus als het specifiek om zoeken gaat, en hier en daar nog wat gespecialiseerde instructies. Een module hier en een dagje daar; heel samenhangend is dat niet. Misschien goed voor levenslang leren, maar het voelt wat onbevredigend als je in een vak moet beginnen. En VOGIN en GO worden ook niet bepaald platgelopen door hordes cursisten.

Maar misschien is zo'n complete opleiding niet altijd meer nodig. Met een goede (andere) vooropleiding is veel benodigde extra kennis bijeen te garen via incidentele cursussen, via congressen, via discussies op het web, via webinars en via kennisuitwisseling met collega's. Zelfstandig kennis opdoen en bijblijven zijn veel makkelijker geworden dan dertig jaar geleden. En laat ik nu ook maar bekennen dat ik zelf nooit een formele opleiding in het informatievak heb genoten; ook dertig jaar geleden al niet. Toch is het met mij nog goed gekomen (geloof ik). Zelfs mag ik nu in Paramaribo anderen bijscholen. De eerste module - over toegankelijk maken van informatie - werd daar enthousiast ontvangen en was uiterst plezierig om te doen. En als u dit leest begin ik net aan de tweede, over "zoeken en vinden". De harde kern van het informatievak hebben ze dan in elk geval gehad.

Bentoistisch zoeken

Een zoeksysteem in Bento-stijl. Toen ik daar kort geleden voor het eerst van hoorde, zei dat me niets. Trendy eet- en gezondheidsrages die op Pinterest de ronde doen, volg ik kennelijk onvoldoende. Want daar bleek "bento" regelmatig in die context op te duiken, toen ik er maar eens naar Googlede. Het blijkt een Japans woord te zijn voor de daar gebruikte lunchboxjes, met aparte vakjes voor de afzonderlijke hapjes en ingrediënten die je van huis of uit de Bento-shop wilt meenemen.

Wat dat met zoeken te maken heeft? Dat is een heel verhaal. Uit analyse van logfiles van zoeksystemen, blijkt dat gebruikers vaak "verkeerde" zoekvragen stellen. Bijvoorbeeld heel specifieke vragen aan een systeem waar alleen heel andere of heel algemene dingen in zitten, met als gevolg dat gebruikers nul hits krijgen - soms wel voor 90% van de vragen. Maar ja, hoe moeten die ook weten waarin precies gezocht wordt als een zoekvenstertje ze uitnodigt om iets in te tikken.

In principe lijkt dat op te lossen met de huidige discovery tools, waar gewoon alles, uit een heleboel informatiebronnen in één zoekindex wordt gestopt. Aan het probleem dat je dan wel erg verschillendsoortige dingen in één bak stopt en dus ook uit je zoekactie krijgt, is iets te doen met parametrische zoektechnieken. Die laten je een zoekresultaat filteren op basis van aan het materiaal toegekende formele metadata.

Een ernstiger probleem is dat je niet altijd alles in die ene discovery index kunt krijgen. Dan lijk je toch weer te moeten terugvallen op metasearch, waarvan we zo blij waren dat het door een discovery tool was vervangen. Bovendien geeft dat meestal alleen maar meer van hetzelfde soort antwoorden die alleen uit afzonderlijke containers zijn gehaald.

Maar dan komt Bentoistisch zoeken te voorschijn. Dat haalt meer van wat anders uit verschillende systemen. Ook wel een soort metasearch, maar dan gepresenteerd in afzonderlijke hokjes die er net zo uitzien als de vakjes van de bento lunchbakjes. Vandaar. Bovenin elk vakje staat prominent vermeld om wat voor soort i nformatie, gegevens of bronnen dat gaat.

Wie een veel te specifieke onderwerpsvraag stelt aan een systeem met alleen algemene metadata van beschikbare databases, krijgt dan automatisch te zien dat het bakje van de Discovery Tool - waar je de vraag niet gesteld had - wel antwoorden geeft. En andere, wellicht onbekende bakjes misschien ook wel. En wie aan de Discovery Tool advies vraagt hoe beter te zoeken, krijgt te zien dat de collectie online instructies (bijvoorbeeld in Libguides) veel beter antwoord geeft dan de inhoud van die Discovery Tool. En als daar misschien wel een antwoord inzat, zou dat verdronken zijn in miljoenen andere documenten - net als bij Google.

Technisch gezien dus niet zo veel nieuws onder de zon. Een lepeltje metasearch, met een scheutje parametrisch zoeken. Het nieuwe is vooral de visuele presentatie, alsof je naar uitnodigende bento-bakjes kijkt.

Digitale datakaalslag

Vorige maand, bij het afscheid van Bas Savenije als directeur van de KB, was informatiefilosoof Luciano Floridi de inhoudelijke feestredenaar. Van zijn moeilijke verhaal kan ik niet heel veel meer reproduceren. Maar één onderdeel is me wel goed bijgebleven, omdat dat voortborduurde op een stokpaardje van mijzelf. De onbeheersbare groei van de hoeveelheid digitale informatie die we produceren. Bij de cijfers die hij gaf, benadrukte ook Floridi met bits en bytes te rekenen en niet met meer inhoudelijke eenheden content.

Nieuw voor mij was Floridi's bewering dat er al een paar jaar niet genoeg opslagcapaciteit meer is om al die data blijvend te bewaren. Al baseerde Floridi die uitspraak op een wat rommelig oud grafiekje van marktonderzoeksbureau IDC, toch zal dat zeker waar zijn - of anders wel snel waar worden. Zelf had ik dat eigenlijk ook kunnen bedenken, want als de hoeveelheid data die we produceren elk jaar verdubbelt, en halvering van zowel prijs als benodigd (fysiek) volume per terabyte twee jaar duurt, dan kan een wiskundige wel uitleggen dat dat mis gaat. Want uitgaven en totaal fysiek volume voor al die opslagmedia kunnen ook niet eeuwig exponentieel blijven groeien om dat verschil op te vangen. Zo had ik zelf ooit geëxtrapoleerd dat we in 2110 alle atomen op aarde nodig zouden hebben om onze data van dat jaar te registreren. Maar een eeuw eerder zitten we dus al in de problemen.

Floridi wilde met zijn verhaal vooral benadrukken dat al lang niet meer alles bewaard kan blijven - hoe goed Archive.org ook zijn best doet. Maar vooral dat het nogal willekeurig is welk deel van ons digitaal geheugen onherroepelijk verloren gaat. Dat dat niet alleen om poezenfoto's gaat, of om scheldpartijen van reaguurders, blijkt uit een recente klacht van een Amerikaanse journalist/blogger die gemerkt had dat ineens meer dan 2000 van zijn blogposts van de laatste vijf jaar verdwenen waren, omdat de betreffende websites gewist waren. Zelf had ik ook al ontdekt dat een heleboel van mijn eigen spullen verdwenen zijn, ook oude blogposts op onze onvolprezen IP-site. Voor die willekeurige digitale kaalslag had Floridi intussen wel een oplossing bedacht: curation! Hij vindt het een mooie nieuwe taak voor bibliothecarissen om beredeneerde keuzes te maken wat bewaard moet blijven.

Floridi's zorgen over wat onherroepelijk verloren gaat, riepen bij mij overigens de vraag op of dat in het predigitale tijdperk niet minstens zo erg het geval was. Veel informatie zat toen alleen in de hoofden van mensen. En van die hersens bestond ook geen back-up als iemand dood ging. Zoals waarschijnlijk zovelen, heb ik zelf vaak betreurd ouders en grootouders niet veel meer te hebben uitgehoord over hun belevenissen - zelfs nog tot WO-I terug - voordat het onherroepelijk te laat was. Maar al had ik dat wel gedaan, over dertig jaar zouden die herinneringen samen met mijn eigen brein toch nog gewist zijn.

Mobiel zoeken - zegen of vloeken

Het is niets nieuws als ik vertel dat een steeds groter deel van de zoekopdrachten die zoekmachines binnenkrijgen afkomstig is van mobiele apparaten. Het aandeel van mobieltjes schijnt begin dit jaar de PC's zelfs al te hebben ingehaald. Niet zo gek, want er zijn intussen ook meer dan twee maal zo veel mobiele apparaten als PC's. Net zo goed als wereldwijd meer mensen een mobieltje hebben dan een WC (maar dat terzijde).

Veel van die "queries" zijn ook niet het soort zoekacties dat wij als informatieprofessionals plegen te doen - als we aan het werk zijn althans. Veel zoekvolume betreft eenvoudige *"local search"*, wat je natuurlijk vooral doet als je met een mobieltje in je hand ergens staat en ter plekke iets lokaals wilt weten. Veel van dat zoeken komt ook uit Apps, in plaats dat een browser wordt gebruikt.

Intussen lijkt deze ontwikkeling ook gevolgen te krijgen voor onze professionele informatie-processen. En zelfs in meer opzichten. In de eerste plaats zal het u wel al zijn opgevallen dat steeds meer websites hun tekst - ook op een gewone PC - in heel grote letters presenteren. Heel handig dat je achter de PC nu geen leesbril meer hoeft op te zetten. Maar het is natuurlijk vooral ingegeven door de behoefte om met hetzelfde interface ook mobiele gebruikers met hun kleine schermpjes te kunnen bedienen. Of je dat ook al *"adaptive web design"* mag noemen is natuurlijk maar de vraag.

Maar daarnaast lijkt het ook nog wat consequenties te hebben die minder positief zijn. Regelmatig ontdek ik dat op een op deze manier vernieuwde site bepaalde keuzes of functies niet meer worden aangeboden, omdat ze - kennelijk - niet meer in die vereenvoudigde mobiel-georiënteerde lay-out passen. En zelfs als er nog wel een apart scherm met PC-opmaak is, blijkt zo'n extraatje daar voor het gemak soms ook maar vast verwijderd.

Een andere ontwikkeling heeft misschien ook wel voor een deel te maken met de opkomst van mobiel: de trend dat zoekmachines op queries reageren met "antwoorden in plaats van 10 blauwe links". Zo'n antwoord past makkelijker op het mobiele schermpje dan een lijst met tien blauwe links. En voor gebruikers van mobieltjes zijn dergelijke concrete antwoorden vaak ook voldoende. Voor onze professionele zoekvragen in de meeste gevallen niet.

Of wij als informatieprofessionals gediend zijn van deze ontwikkelingen, is voor grote spelers op het web natuurlijk niet interessant. Voor Google zijn wij een *quantité négligeable*. Voor hen komt het geld - indirect - van de mobiele zoekers die lokaal iets willen consumeren of kopen. Als wij ons niet willen voegen naar wat Google c.s. goed dunkt, kunnen we hooguit uitwijken naar de alternatieven die er nog altijd zijn. Bij sites binnen onze eigen omgeving moeten we in deze zaken dan wel aansturing geven. Alleen maar vloeken hierover helpt niet.

Twee eeuwen Boole

Dit jaar wordt hier en daar groots gevierd dat George Boole 200 jaar geleden werd geboren. Dat "daar" is nog niet eens zo zeer in zijn Engelse geboortestadje Lincoln, maar vooral in het Ierse Cork. Daar heeft Boole met zijn baanbrekende werk de lokale universiteit op de kaart gezet, in een tijd dat daar nog geen Times Higher Education university rankings voor nodig waren. Vanaf 1849 was hij daar hoogleraar. Boole was inderdaad de man waar onze huidige Booleaanse operatoren naar genoemd zijn. Hoewel hij nog veel meer wiskundigs op zijn geweten heeft, zoals ontdekkingen op het gebied van differentiaal-vergelijkingen en de waarschijnlijkheidsrekening, kennen wij hem natuurlijk vooral van zijn aanzet tot de verzamelingenleer. Onze AND- en OR-relaties zijn voortgekomen uit het doorsnijden en verenigen van die verzamelingen.

Maar juist op die Booleaanse combinaties wordt door de grote geesten van de information retrieval tegenwoordig nogal neergekeken. In 1999 op de allereerste IP-lezing - de voorloper van de huidige VOGIN-IP-lezingen - hadden we Keith van Rijsbergen als keynote spreker uitgenodigd. Deze van oorsprong Nederlandse informaticus was (en is) de goeroe van de theorie van de information retrieval. In een interview dat we hem na zijn lezing afnamen, sprak hij de legendarische woorden "Het Booleaanse model is natuurlijk vreselijk".

Op zich niet zo gek dat hij die mening verkondigde, want hij komt uit de Cambridge school van Karen Spärck-Jones, bedenkster en boegbeeld van de probabilistische zoekmethode. Die techniek achtte hij dan ook verreweg superieur aan dat domme Booleaanse combineren.

Al tijdens het interview begon ik Van Rijsbergen voorzichtig tegen te spreken. Want zestien jaar geleden, was onze zoekpraktijk nog bijna helemaal gebaseerd op AND, OR en NOT. Erg veel indruk maakte mijn tegensputteren uiteraard niet. Toch is het in de praktijk niet zo zwart-wit als Van Rijsbergen het toen voorstelde - nu nog steeds niet. Wij plegen nog altijd vrolijk Booleaanse combinaties te gebruiken, ook bij Google, soms zelfs zonder dat we ons daarvan bewust zijn. Maar tegelijkertijd is de ranking van zo verkregen resultaten wel gebaseerd op die superieur geachte probabilistische technieken - nu bij Google, maar zestien jaar geleden bij AltaVista en Lycos ook al.

Ook in Cork schamen ze zich allerminst voor de bijdrage die George Boole aan de verzamelingenleer heeft geleverd. Breed wordt daar uitgemeten dat zijn wiskundige ontdekkingen aan de basis hebben gestaan van ons complete digitale tijdperk. En inderdaad, in relationele databases wil je liever niet dat resultaten van SQL-queries op probabilistische wijze tot stand komen. Maar daarnaast is het een geluk dat George Boole toch ook nog iets aan waarschijnlijkheidsrekening heeft gedaan. Zo kunnen we proberen om zelfs dat probabilistische zoekmodel nog een heel klein beetje op zijn conto te schrijven.

* Keith van Rijsbergen: ‘Het Booleaanse model is natuurlijk vreselijk’ (IP, 1999) - http://sieverts.pbworks.com/f/Sieverts%2011.99.pdf

De mailtjes van Hillary

In de prachtige Art Deco bioscoop Tuschinski bezocht ik vorige maand een "kaskraker" over zoeken en over *recall* en *precision*. Wie had dat ooit kunnen denken? Maar het was dan ook wel een besloten voorstelling. Hoofdthema van deze ook door IP aangekondigde Amerikaanse documentaire "The decade of Discovery", was de problematiek van de terugvindbaarheid van Amerikaanse interne overheidsinformatie. In de VS zitten daar nog veel meer juridische regels en haken en ogen aan, dan wij hier gewend zijn. Ivo Opstelten had er nog een boel van kunnen leren.

Tot mijn verrassing ging het ineens uitgebreid over een legendarisch artikel van Blair & Maron uit 1985. Dat artikel over recall en precisie was in dertig jaar al geleidelijk uit mijn herinnering weggezakt. Het beschrijft een onderzoek naar het zoekgedrag van advocaten op een groot Amerikaans advocatenkantoor. Die advocaten schatten zelf in dat ze in hun interne full-text zoeksysteem een recall van minstens 75% haalden. Uit dat onderzoek bleek dat in de praktijk nog geen 20% te zijn. Met hun zoekacties misten ze domweg meer dan 80% van de voor hun zaken relevante documenten. In de Amerikaanse rechtspraktijk was dat volsterkt onacceptabel. Simpele full-text zoeksystemen bleken dus niet de destijds verwachte panacee voor onze informatie-problemen.

Sponsor ZyLab had niet alleen de film naar Nederland gehaald en een zaal in Tuschinski afgehuurd, maar ook een van de hoofdpersonen uit de film laten overkomen om een inleiding te geven.

Deze Jason Baron was voormalig directeur van de "Litigation-support" afdeling bij de *U.S. National Archives and Records Administration* die alle overheidsinformatie moet archiveren. Eerder dit jaar was hij ook volop in het (Amerikaanse) nieuws geweest met zijn commentaar op Hillary-gate, de affaire rond de mailtjes van Hillary Clinton. Als minister van buitenlandse zaken had zij ook voor staatszaken ten onrechte haar privémail gebruikt.

Binnen het Amerikaanse rechtssysteem is het volledig kunnen terugvinden van overheidsinformatie, inclusief alle mailcorrespondentie, een heilige graal geworden, zowel ten behoeve van een slagvaardige overheid, als voor het recht van burgers op toegang tot alle digitale informatie. Om Blair & Maron problemen te voorkomen hebben de National Archives strenge regels opgesteld voor een open XML-formaat om alles, ook de mail-correspondentie, te structureren en voor de daarbij te gebruiken zoeksoftware. Dat ZyLab daar een belangrijke rol in speelt, kregen we maar tussen neus en lippen door te horen.

Dat die structurering grootschalig aangepakt moet worden, werd geïllustreerd door cijfers over de email-correspondentie van het Witte Huis. Onder Bill Clinton - de "Clinton-administration" - waren hooguit 30 miljoen mailtjes geproduceerd die de moeite waard waren gearchiveerd te worden, van de Bush-administration waren dat er al 200 miljoen en voor de Obama-administration wordt op ruim meer dan een miljard gerekend. Hoeveel zullen dat er onder Hillary dan wel niet worden? Opdat die allemaal terugvindbaar blijven, moet zij dan natuurlijk wel de goede mail-server gaan gebruiken.

* The Decade of Discovery - https://vimeo.com/106504935

* D.C. Blair & M.E. Marron / An evaluation of retrieval effectiveness for a full-text document retrieval system (1985) - http://yunus.hacettepe.edu.tr/~tonta/courses/spring2008/bby703/blair-maron-evaluation.pdf

De waarheid en niets dan de waarheid

Moet Google de waarheid vertellen? Dat is een vraag die het laatste jaar in allerlei varianten opduikt. Het meest recent stelde The Atlantic die vraag aan de orde onder de expliciete titel "*Should Google Always Tell the Truth?*". Eind vorig jaar werd een soortgelijke vraag ook al op Mainstreet.com gesteld: "*Is Google Responsible for Delivering Accurate - And Truthful - Search Results?*"

Vanuit technisch zoekmachineperspectief was ik geneigd simplistisch te stellen dat zoekmachines alleen maar indexeren en ophalen wat elders geschreven is, zonder zelf verantwoordelijk te zijn voor wat dat is. Bibliotheken zijn toch ook niet aansprakelijk dat alles wat in hun boeken staat, waar moet zijn. Maar misschien zijn zoekmachines toch wat minder neutrale doorgeefluiken. Die doen achter de schermen van alles met die zoekresultaten - en zeker Google. Ze ranken de resultaten en steeds vaker geven ze zelfs feitelijke antwoorden. In hun ranking zitten allerlei factoren die veel verder gaan dan alleen maar lokatie en frequentie van woorden. En ook veel verder dan alleen maar de linkpatronen die in pagerank-achtige algoritmen verwerkt worden. Google heeft onlangs zelfs een *truth algorithm* aangekondigd waarbij de ranking van een webpagina mede afhangt van het waarheidsgehalte dat Google daar automatisch aan toekent op basis van vergelijking met de inhoud van andere pagina's.

Niet zo gek dus om de ethische kant van zoekalgoritmes ter discussie te stellen. "*Should search algorithms be moral?*". Die vraag werd door Quartz (qz.com) voorgelegd aan Google's huisfilosoof, de hier in ander verband al eens aangehaalde Luciano Floridi. Hij vindt in elk geval dat Google niet moet gaan voorschrijven wat waar is. Als de eerste zoekresultaten over het uitsterven van de dinosauriërs toevallig allemaal verwijzen naar verklaringen op basis van de bijbel, dan zij dat maar zo. Toch komt het volgens mij anders te liggen als niet alleen de eerste tien zoekresultaten naar de bijbel verwijzen, maar Google's feitelijke antwoord dat ook doet.

Dat die rankingfactoren ondoorzichtig zijn, riep ook - opnieuw - de vraag op in hoeverre zoekmachines zelfs verkiezingresultaten kunnen beïnvloeden. Met Amerikaanse verkiezingen op komst is dat een terugkerend thema. Zou Google door bewust manipuleren van de ranking van zoekresultaten voor kandidaten, daarmee ook de kiezers kunnen manipuleren? Wetenschappers beweren van wel. Ook dat illustreert dat die vraag naar de ethiek zo gek nog niet is. Bij het Right-to-be-Forgotten speelt de waarheidsvraag natuurlijk ook een rol. Alleen eigenlijk andersom. Die regelgeving maakt dat Google de waarheid soms juist niet meer mag vertellen, wanneer dat bepaalde mensen slecht uitkomt. Waar gaat de ethiek dan voor? Voor privacy of voor de waarheid?

Naar aanleiding van het Atlantic-artikel concludeerde Phil Bradley in zijn weblog al dat het natuurlijk altijd het beste is om dingen aan een bibliothecaris te vragen. Die heeft methoden om onzin van waarheid te onderscheiden. Zou het echt?

* Should Google Always Tell the Truth? (The Atlantic, 2015) - https://www.theatlantic.com/technology/archive/2015/07/should-google-always-tell-the-truth/397697/
* Is Google Responsible for Delivering Accurate - And Truthful - Search Results? (The Street, 2014) - https://www.thestreet.com/story/12929379/1/is-google-responsible-for-delivering-accurate--and-truthful--search-results.html
* Should search algorithms be moral? A conversation with Google's in-house philosopher (Quartz, 2015) - https://qz.com/451051/should-search-algorithms-be-moral-a-conversation-with-googles-in-house-philosopher/

#100

Eigenlijk zou ik een feestje moeten geven, want dit is mijn 100ste column voor IP. Hoewel dat ook weer niet zoveel voorstelt, want NRC-columnist S.Montag (Henk Hofland) is nu al aan aflevering 1820 toe. Maar mij mag je natuurlijk sowieso niet vergellijken met zo'n oercolumnist. Mijn eerste column verscheen in 1997, in de eerste jaargang van IP. Toen verzorgden redactieleden de column nog beurtelings, afhankelijk wie inspiratie had. Daardoor glipten er wel eens columnloze nummers tussendoor. In die tijd waren er jaren waarin ik maar één column schreef en andere jaren met wel vier of vijf. In 2008 veranderde dat. De toenmalige hoofdredacteur vond dat lezers van de column daarin een vaste "smoel" moesten herkennen. En die smoel werd ik - sindsdien ook met foto erboven. Intussen heeft IP nu zelfs al drie vaste smoelen.

In mijn zevende column liet ik ooit weten geen laptop op vakantie mee te nemen. Dat is sindsdien wel veranderd. Als vast columnist moest ik ook wel columns vanuit mijn tentje versturen, of vanaf een leuk terrasje. Zo ook deze.

Waar ik het die honderd keer over gehad heb, moet u zelf maar eens bekijken. De tekst van de voorgaande 99 columns heeft onderstaande woordenwolk opgeleverd. "Google", "informatie" en "zoeken" springen er uiteraard uit. En ook een positieve houding, want "wel" komt het allermeest voor. Jammer dat zo'n woordenwolk geen recht doet aan incidentele gekke onderwerpen die aan de orde kwamen, zoals ontploffende CD-ROMs of het Domesday Book. Daarvoor moet u alle columns zelf nog maar eens doorlezen. Ze staan bij elkaar op mijn website (en nu in dit document).

Misschien is "honderd keer" wel een mooi moment om er een punt achter te zetten. Binnenkort zal ik dat voorzichtig bij de directie aankaarten. IP houdt dan nog altijd twee andere "smoelen" over. En die 1820 van Montag haal ik toch niet meer in.

Twee van de tweehonderd

Voor mijn doen vroeg wakker geworden, had ik al zachtjes de radio aangezet. Tussen sportroddels door kwam plotseling het bericht langs dat het 9de eeuwse Psalter van de Utrechtse UB was toegelaten tot Unesco's prestigieuze "Memory of the World" register. De Utrechtse historicus Marco Mostert lichtte de vroege luisteraars van NPO Radio1 daar uitgebreid over in. Daarbij meldde hij uiteraard ook dat het rijk geïllustreerde handschrift een maand lang tentoongesteld wordt in het Catharijneconvent. Toch was ik wat teleurgesteld over wat niet gezegd werd - en wat ik ook in verdere berichten in de media niet ben tegengekomen. Namelijk dat er ook een digitale versie is, die voor iedereen - zelfs zonder museumjaarkaart - op internet toegankelijk is.

Het is natuurlijk heel bijzonder om zo'n uniek stukje cultuurgoed in werkelijkheid te kunnen zien. Alleen mogen toeschouwers natuurlijk niet bladeren, zodat je niet meer dan twee van de tweehonderd bladzijden te zien kunt krijgen. Dat is toch een beetje alsof je naar het Rijksmuseum gaat om alleen een stukje van 30x30 centimeter van de Nachtwacht te kunnen bekijken. Nee, dan de digitale versie van het Psalter. Daarvan kun je wel al die tweehonderd bladzijden op je scherm krijgen. En bovendien kun je daarop heel ver inzoomen. Je ziet dan meer details dan zelfs met een loupe vlak boven het perkamenten exemplaar zichtbaar zouden zijn.

En dan is er in dit geval nog iets bijzonders. De overvloedige levendige tekeningen en de schat aan details daarin hebben allemaal betrekking op fragmenten uit de daarnaast gecalligrafeerde psalmteksten. Het is eigenlijk een soort stripboek avant-la-lettre. Maar wel heel ver "avant" - bijna twaalf eeuwen zijn tijd vooruit. En in de digitale versie zijn al die relaties tussen beeld- en tekst-fragmenten ook allemaal aanwezig als aanklikbare tweerichtinglinks. Ook wie niet zo bijbelvast is, komt zo te weten waar al die plaatjes vol activiteiten betrekking op hebben.

In 1996 was in Utrecht ook al een digitale versie geproduceerd. Toen nog niet op internet - je moest hem op CD-ROM aanschaffen. In de eerste jaargang van IP is destijds uitgebreid beschreven hoe die tot stand gekomen was en welke speciale mogelijkheden de digitale techniek ons bood. In die versie zaten namelijk ook al aanklikbare koppelingen tussen beeldfragmenten en de stukjes psalmtekst die ermee geïllustreerd werden. Een mooi voorbeeld dat digitale versies van ons erfgoed meer mogelijkheden kunnen bieden dan de fysieke - en dat kan dus heel wat verder gaan dan alleen besparingen op wereldwijde reiskosten. Jammer dat die vorm van *Digital Humanities* voor het grote publiek soms wat onderbelicht blijft ten opzichte van vaak overvloedige PR voor de fysieke *experience*.

Noot: Hoewel ik in mijn vorige column aankondigde dat het de laatste zou zijn, heb ik op verzoek van de redactie toegezegd deze jaargang nog af te maken

* The annotated Utrecht Psalter - http://psalter.library.uu.nl/

Many shades of grey

Begin december was ik bij een internationaal congres over grijze literatuur, dat in Amsterdam werd gehouden. Dat leverde reminiscenties op met de jaren '80 (van de vorige eeuw). Toen waren speciale bibliografische databases voor grijze literatuur, zoals NTIS en SIGLE, nodig om een beetje systematisch te kunnen achterhalen of over een onderwerp documenten of rapporten bestonden, die niet in gewone tijdschriften gepubliceerd waren. Nu we alles op internet zetten, is materiaal dat niet in de officiële publicatiekanalen terecht komt, heel vaak gewoon daar te vinden. Daardoor had ik het gevoel dat grijze literatuur niet echt meer bestond. Althans niet als probleem, waaraan je actief iets moet doen. Toch is er nog wel degelijk een gemeenschap die zich hiermee bezighoudt, verenigd rond Greynet en dit jaarlijkse *Grey Literature* congres.

In deze kringen wordt zelfs veel meer als *grey* geclaimd dan vroeger. Ook sociale media, ook niet-tekst materiaal (beeld, video, audio), ook datasets en zelfs in het algemeen open access publicaties. Hoeveel "grijs" er is, hangt helemaal van je definities af. Je komt het hele spectrum tegen, van "niets is nog grijs" tot "(bijna) alles is grijs". Heel wat *shades of grey* dus.

Opmerkelijk genoeg waren er maar weinig Nederlandse deelnemers, terwijl Amsterdam toch de hoofdstad van *Greyland* is.

Niet vanwege het grijze weer begin december, en ook niet omdat het congres er werd gehouden, maar omdat Greynet in de Amsterdamse Javastraat gevestigd is en Dominic Farace daarvandaan het internationale grijze imperium bestiert. DANS, als exponent van de data-tak in het grijze spectrum, was de enige Nederlandse organisatie die officieel vertegenwoordigd was.

Alles als grijs claimen heeft misschien ook wel zijn nadelen. Elk van de deelgebiedjes die ik eerder noemde, heeft ook al zijn eigen kongsi voor kennisuitwisseling en innovatie. En een heleboel Nederlandse organisaties zijn daarin wel degelijk heel actief, of het nu om AV-materiaal, datasets of Open Access gaat. Misschien dat al die *shades of grey* het echte grijs een beetje onzichtbaar maken.

Biedt "grijs" dan helemaal geen toegevoegd waarde meer? Voor een antwoord op die vraag zal de gemeenschap bij zichzelf te rade moeten gaan - iets dat ze trouwens ook al doen. Wat is het unieke eigene van het "echt" grijze materiaal? Nadruk kan bijvoorbeeld liggen op selectie, "curation" en duurzame toegankelijkheid van belangrijk materiaal dat elders nog wordt veronachtzaamd. Greynet richt zich daar overigens nu ook al op. Maar hiervoor geldt eveneens dat er vaak al andere gremia zijn, die zich ermee bezig houden. Grijs - in welke *shade* dan ook - is misschien wel een te non-descripte kleur om blijvend geestdrift te wekken. Misschien moet Dominic maar eens drastisch een meer uitgesproken kleurtje kiezen.

Noot:
1: Het grapje in mijn titel is uiteraard niet origineel; op het congres werd er al uitvoerig op gevarieerd.
2: Dit was nu echt de laatste column van deze grijsaard.