Meer informatie en beter zoeken: ook centraal in 2023

Eric Sieverts

Inhoud:

Bijdrage aan:
De informatiemaatschappij van 2023 - Perspectieven op de nabije toekomst.
Onder redactie van Dr. G.J. van Bussel
Lectoraat Digital Archiving & Compliance, Hogeschool van Amsterdam, 2013
978-90-76176-00-0

1 De huidige informatiemaatschappij

De ontwikkelingen in de informatiemaatschappij worden in deze bijdrage vooral belicht vanuit het perspectief van een informatiespecialist, die de informatie-voorziening binnen een organisatie tot taak heeft. Daarbij zal enige nadruk liggen op wetenschappelijke informatievoorziening en externe informatiebronnen, al zullen besproken trends en voorbeelden zeker niet uitsluitend daartoe beperkt blijven.
Voordat ik in meer detail op een aantal specifieke trends inga, probeer ik in deze inleiding al een aantal belangrijke karakteristieken te schetsen van de huidige informatiemaatschappij.
Vrijwel alle nieuw gegenereerde informatie wordt digitaal aangemaakt en is dus in één of andere vorm digitaal beschikbaar. Met uitzondering misschien nog van boeken, is het merendeel daarvan - al dan niet tegen betaling - via internet beschikbaar. Dat heeft bij veel - vooral jongere - mensen de opvatting doen ontstaan dat wat niet op internet te vinden is, in feite niet bestaat.
Voor de professionele digitale informatievoorziening heeft dat een verschuiving met zich meegebracht van gebruik van uitsluitend secundaire (bibliografische) bronnen naar de primaire bronnen - het direct gebruik maken van de oorspronkelijke full-text versies van publicaties.[1] Dat is zeker mede ingegeven door de ervaringen met webzoekmachines als Google. Die geven geen verwijzingen, maar linken meteen naar de gevonden (primaire) informatie op de betreffende websites. Zo valt het vinden van informatie - anders dan vroeger - meteen samen met de aflevering ervan. Dat wordt ook wel kortweg omschreven als "discovery=delivery". Bij veel gebruikers heeft dat geleid tot een verwachtingspatroon van "Instant satisfaction". Uit een door OCLC uitgevoerd gebruikersonderzoek kwam dan ook de conclusie dat gebruikers steeds minder geneigd zijn nog achter (primaire) informatie aan te gaan als het te omslachtig is om eraan te komen: “If It Is Too Inconvenient, I’m Not Going After It” (Connaway 2011).
Dat verklaart ook het succes - ondanks enkele kinderziekten - van de "discovery tools" die in steeds meer universiteiten en hogescholen worden geïntroduceerd als centrale zoekdienst voor alle digitaal beschikbaar materiaal. De eenvoudige op Google geïnspireerde interfaces maken die systemen laagdrempelig. En de directe doorlinking naar de betreffende primaire bronnen, van welke uitgever die ook afkomstig zijn, maakt de "delivery" voldoende "convenient".
Een andere ontwikkeling, de tendens om wetenschappelijke publicaties als "Open Access" ter beschikking te willen stellen, zou ook in dit licht bezien kunnen worden. Als jijzelf of de organisatie waar je werkt geen abonnement heeft op een tijdschrift waarin iets interessants gepubliceerd is, maar dat niet Open Access is, dan is het namelijk zeer "inconvenient" - namelijk duur en bovendien gedoe met digitale betaling - om toch nog aan die publicatie te komen. De werkelijke drijfveer hierachter is echter meestal een meer principiële, niet nog eens aan commerciële bedrijven te willen betalen, om de resultaten te kunnen lezen van onderzoek dat al met overheidsgeld is uitgevoerd.
Een aantal van de hier geschetste kenmerken van de huidige informatiemaatschappij zal in het volgende hoofdstuk "Trends" nog uitgebreider aan de orde komen.

[1] Mijn gebruik van de begrippen primaire en secundaire informatie is zoals dat in een deel van de bibliotheekwereld gebruikelijk is. Dat gebruik wijkt af van de betekenis die deze begrippen in sommige wetenschappelijke disciplines hebben.

2 Enkele trends

2.1 Trend 1: Beter zoeken en vinden dankzij geautomatiseerde analyses

Het lijkt een trend dat iedereen alleen nog maar free-text zoekt, want dat is immers wat je bij Google doet. Wie gebruikt nog metadata? Het antwoord op die vraag hangt er echter maar vanaf hoe je metadata definieert. Als je daarbij denkt aan handmatig toegekende gecontroleerde ontsluiting, dan zal het antwoord in veel gevallen NEE zijn. En in die zeldzame gevallen dat het JA zou moeten zijn, heeft een gemiddelde gebruiker dat vaak niet in de gaten, omdat hij niet meer bewust met dat gecontroleerde vocabulaire wordt geconfronteerd. Door geautomatiseerde analyse van materiaal wordt echter steeds vaker additionele tekst gegenereerd, waarop dat materiaal kan worden teruggevonden. En die verrijking mag je ook best metadata noemen. Daarbij bestaat nog wel een flink verschil tussen tekst- en non-tekst-materiaal, zowel in de manier waarop geanalyseerd kan worden, als in de noodzaak om dat te doen. Foto's en video's hebben immers meestal helemaal geen "free-text" waarop je ze Google-achtig terug zou kunnen zoeken. Waar automatische analysetechnieken als semantisch gekarakteriseerd kunnen worden, komen ze nog apart bij Trend 2 aan de orde. (Overigens is het onderscheid daartussen niet altijd heel ondubbelzinnig).

2.1.1 Herkennen van onderwerpen in tekst

Er bestaan al vrij lang statistische technieken om de inhoudelijk belangrijkste woorden in tekstdocumenten te identificeren, zoals de tf*idf methode: woorden die in een document vaker voorkomen, maar in weinig andere documenten, zijn karakteristiek voor (de inhoud van) dat document. Op die manier kunnen zogenaamde vingerafdrukken van documenten worden gegenereerd. Moderne zoekmachines kunnen woorden uit de vingerafdruk van een document bijvoorbeeld extra gewicht geven ten behoeve van betere relevantieordening. Een al langer bestaande toepassing van die vingerafdrukken is om systemen daarmee, via machine learning technieken, te trainen welke onderwerpsterm(en) (bijvoorbeeld uit een thesaurus) of welke onderwerpscategorie (bijvoorbeeld uit een taxonomie) aan een document moeten worden toegekend. Dit soort verrijking - en dus in feite metadatering - wordt in steeds meer systemen toegepast, vooral buiten het open web. Bij zogenaamde *unsupervised learning* technieken worden patronen (bijvoorbeeld categorieën) zelfs zonder aanwezigheid van een taxonomie of thesaurus uit de beschikbare ongestructureerde tekst afgeleid. Het leidt er in principe allemaal toe dat de betreffende tekstdocumenten beter gevonden kunnen worden dan alleen maar op basis van ongecoördineerd zoeken in alle in de documenten zelf aanwezige woorden.
Tegenover dergelijke analyses vooraf van individuele documenten of volledige collecties, staan statistische technieken die, nadat al een zoekactie is gedaan, het resultaat daarvan kunnen uitsplitsen of clusteren, op basis van het samen voorkomen van groepen kenmerkende termen. Elke cluster wordt dan geacht een deelonderwerp of specifieke context van de zoekvraag te representeren, zodat de gebruiker de voor hem meest relevante kan kiezen.

2.1.2 Herkennen van onderwerpen in beeldmateriaal

Beeldmateriaal heeft in principe geen "eigen tekst" op de woorden waarvan het teruggevonden kan worden. Dat webzoekmachines toch naar afbeeldingen kunnen zoeken, komt doordat afbeeldingen in webpagina's meestal wel door tekst worden omgeven. Maar helaas heeft die tekst lang niet altijd allemaal betrekking op de nabije afbeelding. Daarom zijn voor preciezer zoeken eigenlijk inhoudelijke metadata nodig. Als die niet zijn toegekend, ook niet in de vorm van usertags, zijn er intussen toch wel wat mogelijkheden. Zoeken op basis van gespecificeerde kleuren of op basis van een al gevonden of zelf geüploade afbeelding gaat steeds beter, zoals diverse zoektools op internet illustreren. Toch blijken in dat laatste geval de kleuren van het voorbeeld vaak nog een belangrijker rol spelen dan de echte vormen daarin.
Met dit soort technieken wordt nog altijd niet het echte onderwerp van een afbeelding herkend. Software die dat doet bestaat wel en wordt ook geleidelijk beter. De groep van Cees Snoek aan de Universiteit van Amsterdam speelt op dit terrein al enkele jaren een vooraanstaande rol (Snoek 2008, Huurnink 2012). Met behulp van machine learning technieken trainen zijn een systeem bijvoorbeeld op een verscheidenheid aan afbeeldingen die een bepaald object weergeven. Op basis van de daaruit afgeleide karakteristieken kunnen nieuwe afbeeldingen, ook als objecten daarop onder andere omstandigheden zijn afgebeeld, met redelijke betrouwbaarheid worden gecategoriseerd. Kort gezegd: na van 50 uiteenlopende plaatjes van schapen te hebben geleerd dat daar een schaap op staat, kan het systeem van een onbekend plaatje bepalen of daar ook een schaap op staat (of een ander object waarop het systeem getraind is). Dit kan vervolgens ook worden toegepast op de bewegende beelden die in een video voorkomen. Probleem van deze techniek is wel dat voor elk afzonderlijk object getraind moet worden. De resultaten hiervan worden intussen wel steeds beter.

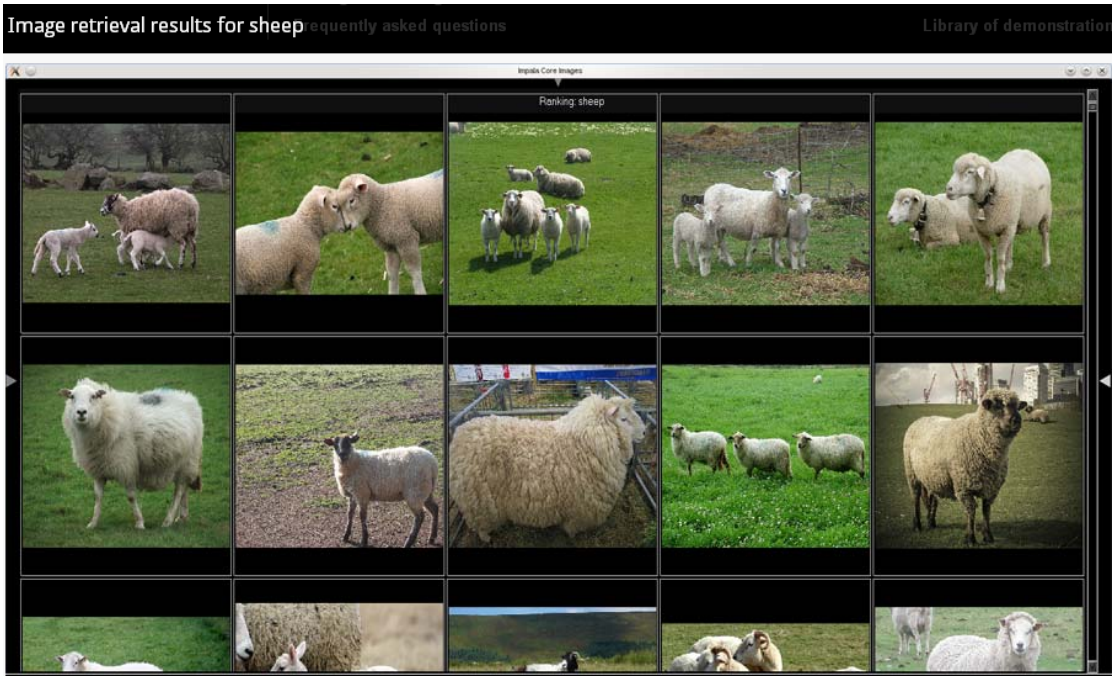

Fig. 2.1 Zoekresultaat uit systeem dat op (onder andere) afbeeldingen van schapen getraind is (Impala systeem - zie overzicht van te herkennen concepten: http://www.euvt.eu/impala/the-concepts-we-can-detect/animals-plants/)

2.1.3 Herkennen van gesproken tekst

Net als aan beeld, is aan digitaal geluidsmateriaal geen directe computerleesbare tekst gekoppeld, waarop een zoekmachine free-text zou kunnen zoeken. Als dat geluid spraak betreft is er echter wel tekst. Om een computer daar iets mee te kunnen laten doen, moeten de woorden in die tekst wel eerst herkend worden. Na de valse start 15 jaar geleden door de firma Lernaut & Hauspie, beginnen er nu wel systemen te komen die in gesproken tekst redelijk kunnen herkennen wat er wordt gezegd. Zoals OCR technieken een afbeelding (scan) van gedrukte of geschreven tekst kunnen omzetten in individuele letters en dus in computerleesbare tekst, zo kunnen audiosignalen ook in woorden worden omgezet. Globaal zijn daarvoor twee soorten toepassingen:

– om gedicteerde tekst of mondelinge opdrachten in een computer in te voeren;
– om audiomateriaal om te zetten in computerleesbare tekst, ten behoeve van presentatie (bijvoorbeeld ondertiteling bij videomateriaal) of om de gesproken tekst full-text doorzoekbaar te maken.

Aanvankelijk moesten systemen voor spraakherkenning getraind worden op de stem van individuele sprekers. Voor apparaten met vaste gebruikers, zoals een dicterende arts of advocaat was dat geen probleem. Gebruik bleef daardoor echter beperkt tot de eerstgenoemde toepassing. Voor het herkennen van door willekeurige sprekers uitgesproken tekst, zoals in radio- en televisieprogramma's, waren die systemen nog ongeschikt. Door training met steeds grotere corpora aan spraakmateriaal is intussen generieke herkenning mogelijk, en is zelfs voor dicteertoepassingen geen training op specifieke gebruikers meer nodig. Die generieke herkenning blijft zelfs niet langer beperkt tot alleen Engels gesproken tekst (nog steeds het grootste gebruiksvolume).

Fig. 2.2 Voorbeeld van Nederlandse spraakherkenning in de Voxalead zoekmachine. De transcriptie laat nog een aantal herkenningsfouten zien.

Een bekende toepassing van de eerste soort is SIRI waarmee iPhone, iPad en iPod van Apple met spraak kunnen worden aangestuurd. Ook Google biedt de mogelijkheid om zoekvragen mondeling in te spreken, waarbij ook Nederlands gesproken tekst onder de meeste omstandigheden zonder al te veel fouten wordt herkend.
Aan de zoekkant breekt spraakherkenning nog maar aarzelend door. Enkele jaren geleden maakte het Amerikaanse EveryZing op internet al Engelstalig videomateriaal full-text doorzoekbaar. Maar dat bleek alleen tijdelijk als demonstratie van hun commercieel te verkopen product. Videozoekmachine Blinkx zegt al enige tijd spraakherkenning te gebruiken. Een meertalig voorbeeld waarbij dat heel duidelijk blijkt, is Voxalead. In VoxaleadNews zijn nieuwsuitzendingen - audio en video - full-text doorzoekbaar. Naast Engels ook in het Frans, Duits, Nederlands, Spaans, Italiaans, Russisch, Arabisch en Chinees. De door de spraakherkenning gegenereerde transcriptie, op basis waarvan gezocht wordt, kan real-time meelopen bij het afspelen van gevonden materiaal. Hoewel je daarbij ziet dat beslist nog niet alle woorden correct worden verstaan, zeker niet in rumoerige situaties, biedt dit al een veelheid aan nuttige zoekingangen op audio en video materiaal. Bijkomend voordeel is dat meteen bij dat fragment gestart kan worden waar de zoekwoorden voorkomen.
Op zang - in feite ook gesproken woord - worden deze technieken nog nauwelijks toegepast. Daarvoor is het ook eigenlijk overbodig, omdat de teksten meestal al als "lyrics" digitaal beschikbaar (en dus doorzoekbaar) zijn. Van muziek kunnen wel al een soort vingerafdrukken worden gemaakt, op basis waarvan bijvoorbeeld Shazam een met een smartphone opgepikt muziekfragment op naam kan brengen.

2.1.4 Analyse van sociaal netwerk en sociale media

Sociale netwerken die op internet vorm krijgt in de toonaangevende systemen (Facebook, Twitter, Google+, LinkedIn, Reddit, Goodreads, ...) spelen steeds sterker ook een rol bij gewone zoekmachines. Activiteiten in die sociale netwerken worden geanalyseerd, zodat datgene wat anderen binnen ons persoonlijk netwerk hebben gedaan, bekeken, gelezen, gemeld of geapprecieerd, door zoekmachines kan worden meegewogen in de beoordeling van de relevantie van wat op onze zoekvragen wordt gevonden. En zelfs zonder te zoeken, ontvangen we al recommandaties die hierop zijn gebaseerd. Overigens draagt dit er wel toe bij dat gebruikers steeds meer binnen hun zogenaamde "filter bubble" blijven en geen - of in elk geval minder - informatie te zien krijgen vanuit andere invalshoeken (Pariser 2011).
Ook kan rechtstreek gezocht worden binnen die "social graph", het netwerk van personen, berichten, objecten en gebeurtenissen die binnen die sociale media aan elkaar gerelateerd zijn. Alles wat Facebook van zijn miljard gebruikers registreert vormt ook zo'n "graph". De onlangs in Amerika beschikbaar gekomen Facebook Graph Search biedt daarin zeer gestructureerde zoekmogelijkheden (Starr 2013).
Een ander soort analyse die op sociale media ingang vindt, is zogenaamde sentiment analysis. Op grond van tekst en andere karakteristieken in de berichten die we achterlaten (ook al zijn het maar de maximaal 140 karakters van een tweet) kan automatisch worden bepaald of daarin positief of negatief geoordeeld wordt. Ook hierbij worden machine learning technieken ingezet (zij het dat ook wel eens simplistischer naar de inhoud van berichten wordt gekeken, zoals ☺ vs. ☹). Dit soort technieken draagt er mede toe bij dat sociale media als bron voor big data kunnen worden geëxploiteerd (zie 2.3.5).

2.2 Trend 2: Beter zoeken en vinden dankzij semantische technieken

Semantisch zoeken is het nieuwe paradigma waarop veel zoekmachineontwikkelingen zich richten - zeker voor het web. Dat "semantisch" houdt in dat zoek- en analyse-technieken worden toegepast om de computer iets te weten te laten komen over de betekenis van woorden in zowel zoekvragen als gevonden documenten. Steeds meer zoektools afficheren zich ook al als semantisch, of dat nu terecht is of niet. In de praktijk blijkt semantisch zoeken namelijk een soort containerbegrip waar men allerlei doelstellingen en technieken onder laat vallen (Starr 2012). Doelen en methoden die vaak genoemd worden in overzichten van wat semantisch zoeken inhoudt, vallen globaal in drie categorieën:

- inschatten van doel of context van zoekvragen,
- vooraf (bij indexeren) of achteraf (na een zoekactie) analyseren van tekst om betekenis van woorden af te leiden,
- automatisch aanpassen van zoekvragen op basis van de betekenis van de oorspronkelijk gebruikte zoekwoorden.

2.2.1 Doel en context van zoekvragen

Sommige zoekmachines proberen een inschatting te maken van het meest waarschijnlijke doel waarvoor een zoekvraag wordt gesteld (Landry 2013). Daartoe kan een zoekmachine diverse soorten indicaties gebruiken, zoals gegevens over de locatie van de gebruiker, eerder zoekgedrag van die gebruiker en de aard en formulering van de zoekvraag. De te gebruiken locatie-informatie kan zich beperken tot de landenversie van de zoekmachine die wordt gebruikt. Nu voor het stellen van zoekvragen steeds vaker mobiele apparatuur wordt gebruikt, kan het ook om veel gedetailleerder informatie gaan, omdat de locatie van een gebruiker zo heel precies bekend kan zijn. De zoekvraag "pizza" kan dan geïnterpreteerd worden als de wens er een te gaan eten en niet als een zoekactie naar een recept of naar documenten over de geschiedenis van de pizza. Informatie over restaurants in de onmiddellijke nabijheid van de zoeker is dan het waarschijnlijk gewenste zoekresultaat.
Ook eerder zoekgedrag van de gebruiker en eerder geraadpleegde websites kunnen een indicatie geven van de waarschijnlijke context waarbinnen een vraag gesteld wordt. Dat zorgt voor personalisatie van zoekresultaten, maar brengt ook het gevaar met zich mee dat gebruikers steeds binnen hun eigen "filter bubble" blijven en geen informatie vanuit andere gezichtspunten meer te zien krijgen (Pariser 2011).
Bij wat uitgebreider geformuleerde zoekvragen kan door natuurlijke taaltechnieken getracht worden de aard van de zoekvraag te achterhalen. Wie "geboortejaar Beethoven" intikt zal waarschijnlijk alleen een jaartal willen weten en geen behoefte hebben aan een uitgebreide lijst zoekresultaten. Bij de door Google in 2012 geïntroduceerde Knowledge Graph (Starr 2013) gebeurt dit al. Zelfs bij kortere vragen naar alleen een persoonsnaam, stad, land of bedrijf, worden vaak meteen al feitelijke gegevens getoond, overigens nog wel naast een lijst gewone zoekresultaten. Bij dubbelzinnig woorden of namen worden dan soms gegevens over de verschillende betekenissen getoond. De zoekmachine WolframAlpha, die zichzelf een "computational knowledge engine" noemt, komt zelfs uitsluitend met feitelijke gegevens op basis van inhoudelijke interpretatie van de zoekvraag. Zo is er een trend dat zoekmachines in feite tot antwoordmachines evolueren.

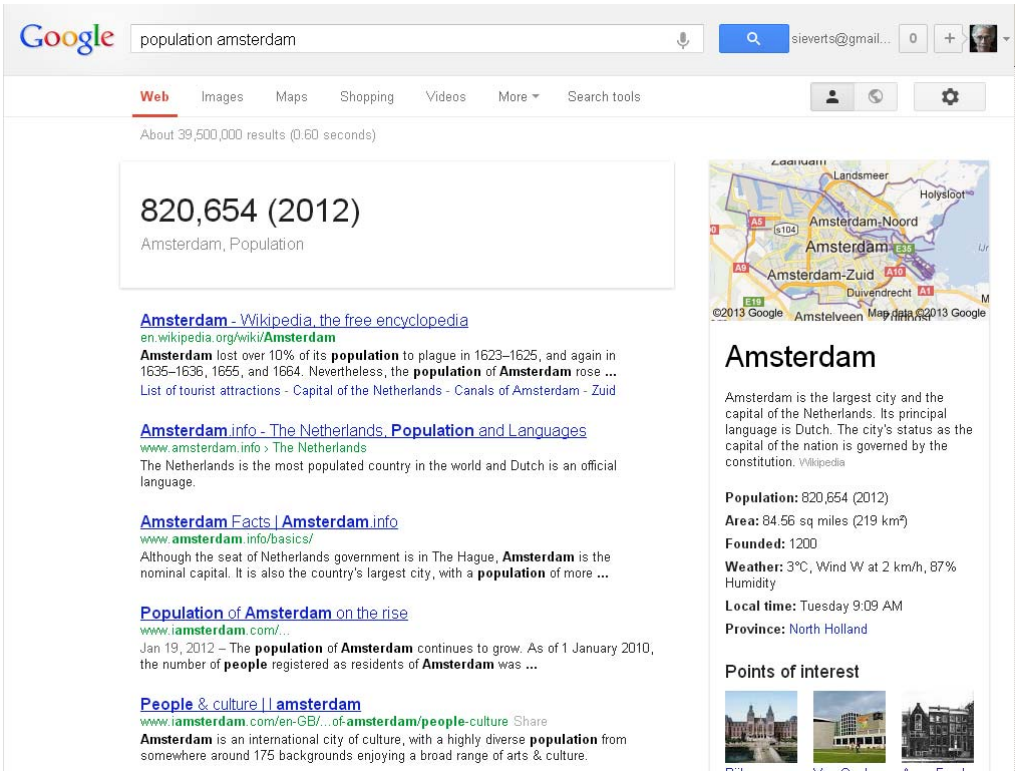

Fig. 2.3 Voorbeeld van feitelijke gegevens die Google presenteert door interpretatie van de bedoeling van een zoekvraag.

2.2.2 Analyse van informatie

In tekstdocumenten kunnen concepten, woorden of namen herkend worden als behorend tot een bepaalde categorie, zoals plaatsnamen, namen van personen, producten of bedrijven, gebeurtenissen en dergelijke. Er zijn systemen waarbij dat herkennen vooraf gebeurt, bij het indexeren van de teksten, zodat gerichte zoek- en filtermogelijkheden kunnen worden aangeboden. Het kan ook achteraf met resultaten van een al uitgevoerde zoekopdracht, met als doel de gebruiker een duidelijker beeld te bieden van wat er is gevonden of om de resultaten te kunnen uitsplitsen. De herkenning van dergelijke specifieke entiteiten kan automatisch plaats vinden op basis van beschikbare lijsten van dergelijke entiteiten, vaak wel in combinatie met taalkundige analyse. Het gebeurt ook steeds vaker op basis van vooraf al in de tekst aangebrachte coderingen, een soort interne (of "embedded") metadata.
Voor metadata die in de HTML-code van webpagina's verwerkt worden, zijn op RDF (Resource Description Framework) gebaseerde methoden ontwikkeld. Op basis daarvan heeft Google voor bepaalde soorten informatie inhoudelijke standaarden vastgelegd voor zinvolle kenmerken van recepten, recensies, personen, producten, organisaties, gebeurtenissen en muziek, zodat bij zoekresultaten zogenaamde "rich snippets" getoond kunnen worden. Recepten van aanbieders die zich aan Google's standaard houden, kunnen zo ook heel gericht gefilterd worden op al dan niet benodigde ingrediënten, bereidingstijd en aantallen calorieën.
Onder de naam Schema.org hebben de drie grote webzoekmachines gezamenlijk ook een meer algemene hiërarchie van standaard begrippen en eigenschappen opgezet, waarmee allerlei soorten informatie in webpagina's van zulke "embedded metadata" kan worden voorzien. Hierbij wordt al van een ontologie gesproken. Ook meer

toepassingsspecifieke ontologieën zijn op het web al in gebruik, zoals de GoodRelations Ontologie voor E-Commerce toepassingen (intussen geïntegreerd in Schema.org) of de Music Ontology voor sites met muziek-informatie. Aanbieders van informatie op het web blijken deze standaarden ook steeds meer toe te passen, omdat men er het belang van inziet dat de aangeboden informatie op deze manier beter gevonden kan worden en in de praktijk vaak ook hoger in resultatenlijsten terecht komt (Landry 2013, Starr 2012). Daarmee is het in feite een onderdeel geworden van de technieken voor Search Engine Optimalisatie.

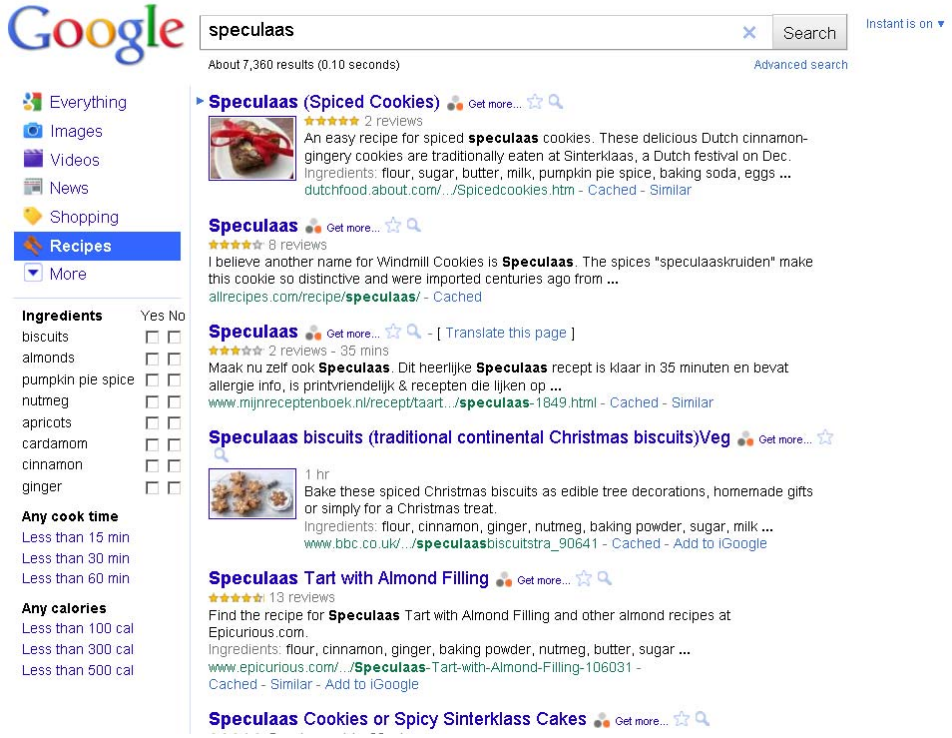

Fig. 2.4 Voorbeeld van extra filtermogelijkheden in Google's receptenzoeker, dankzij aanwezigheid van gestandaardiseerde "embedded metadata" in de doorzochte webpagina's.

2.2.3 Ontologieën en semantisch web

Niet alleen semantisch zoeken, maar ook het "semantisch web" staat volop in de belangstelling. In dat semantisch web zal veel informatie al bij voorbaat met betekenis gelabeld worden, zoals met de zojuist besproken "embedded metadata". Daarbij speelt het begrip ontologie een centrale rol. Formeel wordt een ontologie wel gedefinieerd als "een strikt en uitputtend schema voor een bepaald onderwerpsdomein, meestal in een hiërarchische structuur, die alle relevante grootheden en hun relaties bevat, alsmede de regels waaraan die grootheden en relaties binnen dat domein voldoen". In principe beogen ontologieën op een geformaliseerde en door computers interpreteerbare manier een gedetailleerde beschrijving te geven van (een stukje van) de werkelijkheid. Een ontologie moet daartoe in een formele computertaal beschreven kunnen worden. OWL, de Web Ontology Language is een voorbeeld van zo'n taal.
In toepassingen voor het semantisch web moet een computer echt betekenis van (tekst in) documenten en metadata kunnen afleiden en op basis daarvan kunnen redeneren en gevolgtrekkingen maken. Dat laatste lijkt voorlopig nog wat hoog gegrepen. We

zien dan ook dat ontologieën op dit moment nog vooral worden gebruikt om af te leiden welke begrippen inhoudelijk iets met elkaar te maken hebben en welke woorden min of meer gelijke betekenis hebben, ook als die woorden in verschillende ontologieën voorkomen. Daartoe hoeven niet altijd "echte" ontologieën, volgens de eerdere definitie, beschikbaar te zijn. Ook woordenlijsten, thesauri, taxonomieën, semantische netwerken en concordanties worden in dit kader als ontologie bestempeld. En ook Google's Knowledge Graph, in feite een netwerk van entiteiten met daartussen relaties, mag je wel zo noemen (Gallagher 2012). Dergelijke systemen die termen inhoudelijk met elkaar in verband brengen, ook tussen verschillende systemen, maken het mogelijk zoekvragen beter te interpreteren en ze automatisch aan te passen en uit te breiden met andere inhoudelijk relevante zoekwoorden. Dat geldt ook voor het verbreden of specifieker maken van zoekvragen. Ook kan zo automatisch aanvullende inhoudelijke informatie uit andere bronnen worden opgevraagd, zoals uit Google's Knowledge Graph of uit een linked-data versie van de Wikipedia, de DBpedia.

2.2.4 Linked data en semantisch web

Een belangrijke bron voor gegevens ten behoeve van het semantisch web wordt intussen gevormd door zogenaamde "linked data". Dit zijn feitelijke gegevens die op technisch gestandaardiseerde wijze op internet beschikbaar worden gesteld, vrij voor iedereen te gebruiken, zodat ook wel van "linked open data" wordt gesproken. Deze gegevens zijn afgeleid uit allerlei bestaande systemen en databases. Een belangrijke bron hiervoor is de al genoemde DBpedia, waarvan de beginletters "DB" aangeven dat het om een soort "database"-versie van de Wikipedia gaat. Omdat op zijn minst een deel van de inhoud van elke beschrijving in de Wikipedia, de "infoboxen", gebruik maakt van een vaste structuur, kunnen gegevens daaruit in principe automatisch worden omgevormd tot het standaard format van linked data. Anderzijds is er ook een recent Wikidata-initiatief waarin gevalideerde feitelijke gegevens rechtstreeks worden ingevoerd in linked data vorm, zodat omgekeerd de Wikipedia - en dus ook andere informatiesystemen - van die gecontroleerde feiten gebruik kunnen maken.
Daarnaast zijn er nog vele honderden, meest kleinere collecties van gegevens als linked open data beschikbaar, zoals gegevens uit de Internet Movie Database (Linked MDB), de wereldwijde autorisatielijst van auteursnamen VIAF, collectiegegevens van het Amsterdam Museum, et cetera (Cyganiak 2011). Ook overheden zijn in toenemende mate actief om overheidsinformatie van allerlei aard als (linked) open data beschikbaar te stellen. Vanuit de bibliotheekwereld worden diverse initiatieven ontwikkeld om zoveel mogelijk datasets beschikbaar te stellen (Library 2012).
Diverse ontologieën - in de brede betekenis - worden eveneens in de vorm van linked open data beschikbaar gesteld. In de erfgoedsector is Europeana een systeem dat al veel gebruik maakt van semantisch web technieken en linked open data om gegevens uit diverse collecties te kunnen combineren en gebruikers zo veelzijdiger antwoorden op hun zoekvragen te kunnen bieden.
Zoals uit de hier gegeven voorbeelden blijkt, is deze semantische trend vooral op het web zichtbaar. Dat is een mooie grote proeftuin voor het ontwikkelen en uittesten van semantische zoektechnieken. Maar juist daar spelen ook commerciële motieven mee: wie zich met Search Engine Optimalisatie bezig houdt, wordt al steeds meer gedwongen ook aandacht aan semantiek te besteden.

2.3 Trend 3: Steeds meer, steeds kleiner en steeds goedkoper

Het is uiteraard een open deur te stellen dat er steeds meer informatie geproduceerd wordt - en de laatste decennia natuurlijk vooral digitale informatie. In de volgende paragrafen blijkt dat allerlei parameters al gedurende langere of kortere tijd een exponentiële groei vertonen, soms zelfs met een verdubbelingsperiode van slechts een jaar. Een dergelijke trend is ook te herkennen bij de capaciteit van de media waar deze informatie wordt opgeslagen.

2.3.1 Groei van de wetenschappelijke productie

Voor colleges over "online informatie zoeken" die ik in de periode 1981-1994 bij de Universiteit van Amsterdam heb verzorgd, had ik destijds gegevens verzameld over de productie aan wetenschappelijke literatuur. Dit om duidelijk te maken dat geautomatiseerde technieken voor het zoeken van die informatie intussen onontkoombaar waren. Een belangrijk deel van die gegevens kwam uit publicaties van Derek de Solla Price, de goeroe van de wetenschap van de wetenschap (De Solla Price 1963). Daarnaast had ik ook gegevens gebruikt uit Ulrich's Periodicals Directory en de Gale Directory of Databases.

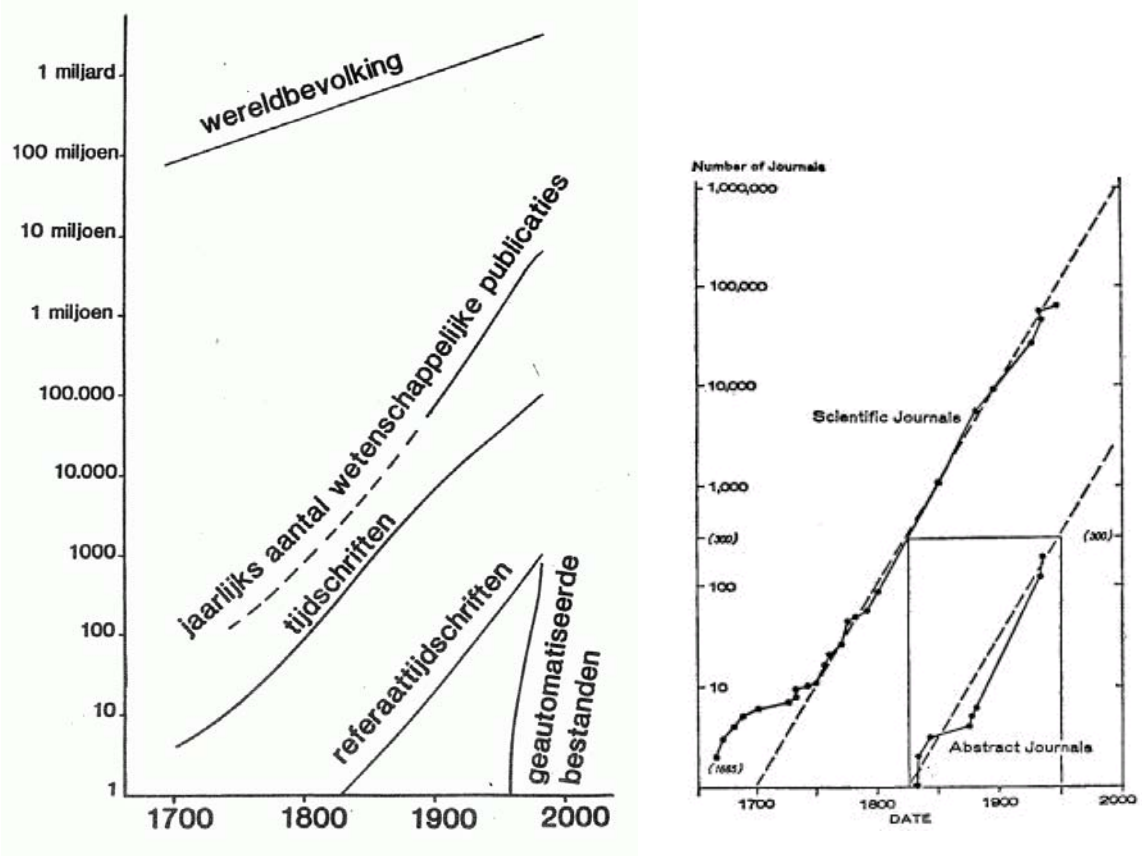

Fig. 2.5 (a) Geschatte groei sinds ca. 1700 van diverse parameters voor de wetenschappelijke informatievoorziening (Sieverts, 1982)
(b) Groei van het aantal wetenschappelijke tijdschriften en het aantal (secundaire) abstracttijdschriften (De Solla Price, 1963)

Dat leverde een grafiek op (Fig. 2.5a) van de geschatte groei van het aantal jaarlijks gepubliceerde wetenschappelijke artikelen, van het aantal wetenschappelijke tijdschriften, alsook van de aantallen referaattijdschriften en (meer recent) online databases. 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 ongeveer iedere 14 jaar. Zoiets is nogal ongebruikelijk. Voordat een verschijnsel met vijf ordes van grootte gegroeid is, treedt meestal verzadiging op, doordat er tekort ontstaat aan voedsel, grondstof of ruimte. Daardoor krijg je een soort afvlakkende S-curve. Ook zonder aanvullend onderzoek, tekende ik deze grafiek elk volgend collegejaar ietsje verder door in de tijd, daarbij wel al rekening houdend met een waarschijnlijk geachte afvlakking.
Sinds begin jaren '90 had ik deze grafiek niet meer gebruikt en dus ook niet verder aangevuld en evenmin aanvullende gegeven gezocht. In de wetenschapsbijlage van NRC-Handelsblad van 12 maart 2011 verscheen echter een artikel dat meldde dat ook de laatste decennia, de 30 jaar sinds 1980, de wetenschappelijke productie nog altijd elke 12 jaar was verdubbeld (Brouwer, 2012). Als de gegevens uit dit artikel (grotendeels gebaseerd op aantallen artikelen die in de Scopus-database zijn opgenomen – zie Fig. 2.6) in mijn oude grafiek worden uitgezet, blijkt dat een opmerkelijke overeenkomst op te leveren. Na – inderdaad – een korte periode van afvlakking, is de groei weer onverminderd doorgegaan. Kennelijk heeft de sterke groei van het aantal wetenschappers in landen als China en India een verdere afvlakking nog wat weten uit te stellen.

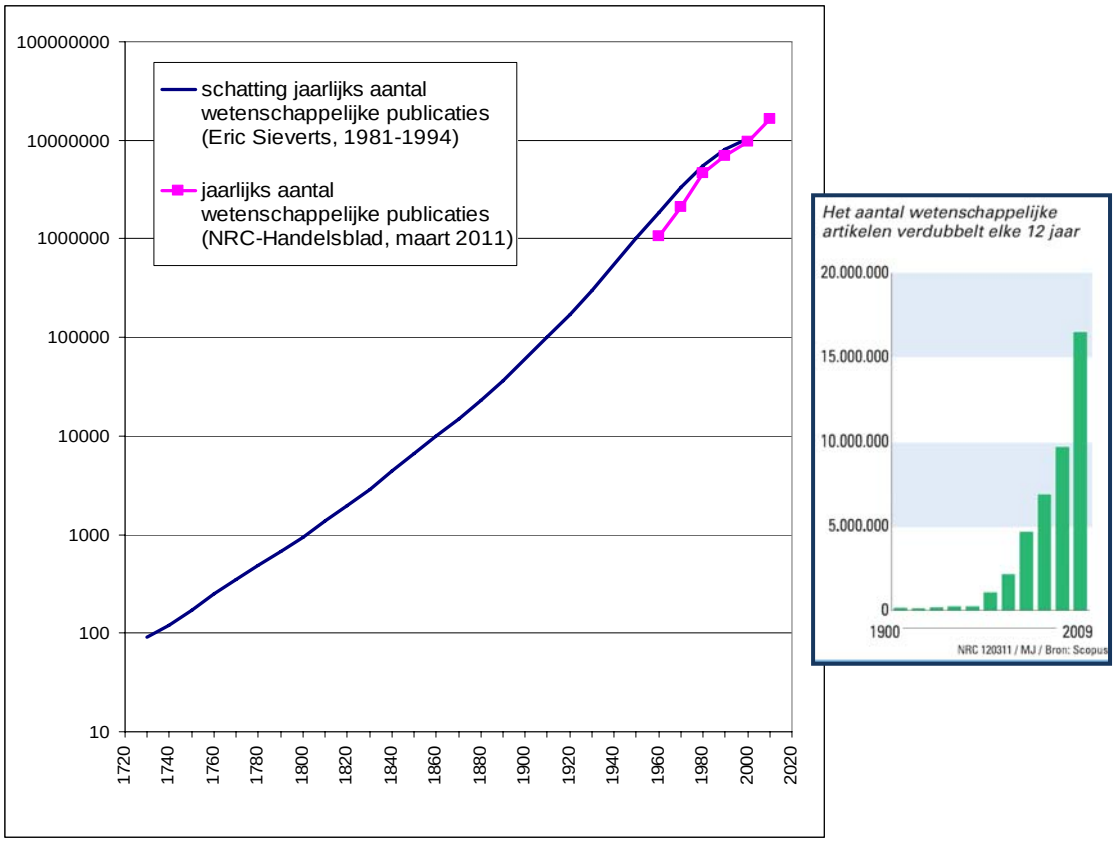

Fig. 2.6 Het jaarlijks aantal wetenschappelijke artikelen uit Fig. 2.5, gecombineerd met de resultaten van een analyse van de Scopus-database (Brouwer, 2012)

2.3.2 Groei van het world wide web

Het opmerkelijke feit doet zich voor dat de grootte van het world wide web één van de grote onbekenden is in het huidige informatielandschap. Het web is over een zo groot aantal knooppunten gedistribueerd, dat daar in de praktijk geen harde cijfers over afgeleid kunnen worden. De groei van het web laat zich nog het makkelijkst afmeten aan de geschiedenis van de zoekmachines. Die geschiedenis begon in feite in 1994 met Lycos, de eerste bruikbare webzoekmachine, gebouwd door een hoogleraar en studenten van Carnegie Mellon University in de VS. Aanvankelijk maakte die 1,5 miljoen webpagina's doorzoekbaar.

In de beginjaren speelde de grootte van webzoekmachines uit concurrentieoogpunt een belangrijke rol. Op de website van SearchEngineWatch (SEW) werd dat voor de belangrijkste zoekmachines bijgehouden in staafdiagrammen en grafieken. Die cijfers waren meestal afkomstig van de zoekmachines zelf. De hoeveelheid zo beschikbare informatie was aanvankelijk veel kleiner dan wat beschikbaar was in databases waartoe online hostorganisaties toen al tegen betaling toegang boden. Maar dat was snel voorbij. Door verdubbeling elk jaar gaven webzoekmachines tien jaar later – puur kwantitatief gezien - al toegang tot een veelvoud van wat betaalde bronnen boden.

Het gegeven van jaarlijkse verdubbeling, komt voor de eerste jaren uit de SEW-gegevens. Maar na een tijdje stopten webzoekmachines met het vermelden van hun groottes, omdat die steeds minder een marketingargument vormden. Zelf heb ik uit nieuwsgierigheid toen enige tijd metingen gedaan. Daartoe liet ik een aantal standaard zoekvragen los op te vergelijken zoekmachines, telde aantallen resultaten op en gebruikte die cijfers - statistisch misschien niet waterdicht - als indicatie voor hun relatieve groottes. Voor absolute cijfers relateerde ik dat aan de laatste nog officieel gerapporteerde groottes. Die cijfers - eerst van SEW, later van mij - zijn in bijgaande grafiek (Fig. 2.7) uitgezet. De vrijwel rechte lijn op de logaritmische schaal illustreert exponentiële groei. Vijftien jaar lang verdubbelde de omvang van de grootste zoekmachine elk jaar. Ten opzichte van de allereerste Lycos-index uit 1994, een factor 100.000 groei. Een groei waar wetenschappelijke publicaties 230 jaar voor nodig hadden, vond op het web in 15 jaar plaats.

Na 2008 loopt die grafiek niet door. De twee belangrijkste redenen daarvoor zijn:

1. Door zoekmachines worden zoekvragen intussen voortdurend op andere manieren ver- en bewerkt, waardoor resultaten niet goed vergelijkbaar meer zijn.
2. Het wordt steeds onduidelijker wat je als afzonderlijke webpagina's moet tellen. Is elke tweet een webpagina? Heb je te maken met een andere pagina als uit een CMS een enkel ander blokje informatie op een pagina worden geplaatst of als verschillende gebruikers op dezelfde site verschillende gepersonaliseerde pagina's met een iets ander dynamisch gegenereerd URL te zien krijgen?

Toch kwam Google in 2012 ineens weer met cijfers. Ze meldden 30 biljoen webadressen (URL's) te kennen. En ook dat alleen al hun index 100 miljoen gigabytes (ofwel 100 Petabyte) is. Dat aantal webadressen is een interessant gegeven om te vergelijken met andere recente pogingen om te meten hoeveel pagina's Google ook werkelijk geïndexeerd heeft - nog wel iets anders dan weten welke URL's bestaan. De cijfers die je daarvoor tegenkomt, liggen meestal vele ordes van grootte lager; op dit moment rond de 100 miljard met minstens een factor twee onzekerheid. Dit grote verschil zal vermoedelijk samenhangen met de pogingen van Google (en andere zoekmachines) om alleen "echte" webpagina's te indexeren. Daarbij worden ondermeer gepersonaliseerde varianten overgeslagen, evenals miljarden spampagina in zogenaamde linkfarms die alleen bedoeld zijn om relevantiescores van echte pagina's kunstmatig te beïnvloeden.

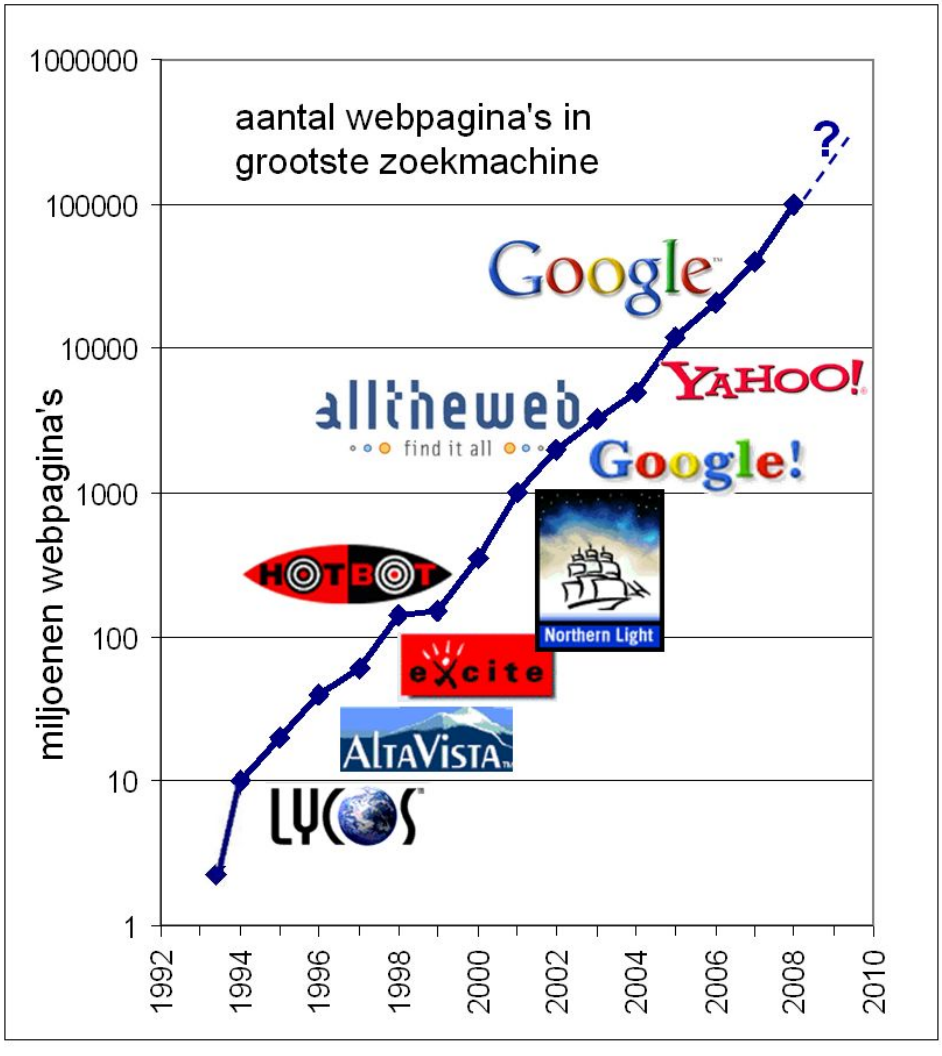

Fig. 2.7 Omvang (in aantal webpagina's) van de in een bepaald jaar grootste webzoekmachine, voor de periode 1994-2008

2.3.3 Groei van de dataproductie

De groei van het web is niet de enige parameter om de groei van onze digitale informatieproductie te meten. Er worden ook al bijna 15 jaar pogingen gedaan om te analyseren hoeveel bytes jaarlijks wereldwijd worden geproduceerd. Die analyses komen er meestal op uit dat die productie ook bijna jaarlijks verdubbelt. Het is goed je te realiseren dat een dergelijke groei inhoudt dat elk volgend jaar evenveel wordt geproduceerd als in alle voorgaande jaren samen!
Een eerste onderzoek hiernaar was dat van Lyman en Varian uit 2000 (Lyman, 2000) waarin werd geconcludeerd dat we in het voorgaande jaar gezamenlijk 1,5 exabyte (miljard gigabyte) hadden geproduceerd. In 2003 was dat onderzoek nog eens ge-updated (Lyman 2003). Daaruit kwam een ruime verdubbeling in 3 jaar. In dat onderzoek zat echter nog een wat onduidelijke vermenging van informatie op papier en born digital materiaal.

Voor een meer recente Infographic over dit onderwerp (Kalakota 2011) werd alleen nog uitgegaan van puur digitaal opgeslagen materiaal. Dat kwam uit op een productie over 2011 van 2 zettabytes (2 x 10^{21} bytes, ofwel 2000 miljard gigabyte). Omgerekend komt dat neer op ruim 300 GB per aardbewoner persoon. Bovendien geeft dit onderzoek aan dat iedereen gemiddeld ook nog eens een zelfde hoeveelheid aan "schaduwgegevens" (back-ups en kopieën) opslaat. Daarbij zorgt mobiele apparatuur voor een jaarlijkse stijging van dataverkeer van 82%, video voor een stijging van 48%. In dat verband nog een paar getallen:

- in 2011 werd per minuut 72 uur videomateriaal geüploaded naar YouTube;
- in 2011 werden 1 biljoen YouTube filmpjes bekeken, ofwel 140 per aardbewoner
- (Youtube 2012)
- in 2012 werden gemiddeld 350 miljoen foto's per dag geüploaded naar Facebook (Sengupta 2013)

Voor 2015 verwachtte Kalakota een totale dataproductie van 8 zettabytes. Vergeleken met de cijfers uit 2000 van Lyman, is dat in 15 jaar een groei met een factor 5000, ofwel een verdubbeling ongeveer elke 14 maanden.
Als we dat voor het gemak even afronden naar een jaarlijkse verdubbeling, dan komt een extrapolatie naar het jaar 2110 uit op een productie van ~10^{51} bits. Aangezien dat aantal in de buurt komt van het aantal atomen waaruit de aarde bestaat, zal het duidelijk zijn dat er voor die tijd zeker grenzen aan de groei zullen zijn - zelfs als bovenstaande berekeningen de groei wat erg overschatten. Voor een kortere termijn, tot 2023 zal deze mate van exponentiële groei echter zeker nog kunnen voortduren, zodat we dan 500 à 1000 maal zoveel zullen produceren als nu.

2.3.4 Capaciteitsgroei en miniaturisatie van opslagmedia

Al die geproduceerde data moeten uiteraard ook ergens worden opgeslagen. De capaciteit en mogelijkheden van dataopslag lijken echter aardig gelijke tred te houden met de juist beschreven productiegroei. Je hoeft maar te kijken naar de schijfgrootte van de opvolgende generaties PC's waarop we werken of naar de capaciteit van de geheugenkaartjes voor onze digitale camera's of smart-phones om daar al enig gevoel voor te krijgen. Visueel wordt dat aardig geïllustreerd in Fig. 2.8, waar we zien hoe een 5 MB harde schijf in 1956 per vliegtuig vervoerd moest worden. De ruim 2 m^3 grote kast moest met een vorkheftruck worden ingeladen. Wanneer we dat vergelijken met de nieuwste geheugenschijven van 60 terabyte in een volume van grof geschat 50 cm^3 of Flash geheugenkaartjes van 2 terabyte in 0,5 cm^3, dan zien we dat in 57 jaar de opslagcapaciteit per volume-eenheid met ongeveer een factor 10^{12} is toegenomen. Uitgaande van een constante exponentiële groei, komt dat neer op een capaciteits-verdubbeling elke 17 maanden. Dat houdt dus net niet helemaal gelijke tred met de verdubbelingstijd van de huidige data-explosie. Dat wordt echter meer dan goed gemaakt door de onvoorstelbare proliferatie van opslagmedia.
Die proliferatie kon uiteraard plaats vinden doordat de kosten voor opslagmedia een soortgelijke trend tonen. Voor de 5 MB schijf van IBM moest nog een jaarlijks lease-bedrag van $35.000 worden betaald. Voor de goedkoopste opslagmedia nadert de prijs op dit moment de magische grens van $10 per terabyte. Die kostendaling is overigens pas echt ingezet na de komst van de PC, begin jaren 1980. Als we daarom alleen naar de laatste 30 jaar kijken, dan heeft in die periode een prijsdaling met een factor 10^7 plaats gevonden (Havard, 2013). Dat komt overeen met een halvering van de prijs elke 16 maanden.

De hier geschetste groei van opslagcapaciteit vindt uiteraard niet alleen plaats op lokaal niveau in de apparatuur die we zelf bij de hand hebben. Steeds meer data worden opgeslagen bij via internet bereikbare gespecialiseerde diensten die door hun schaalgrootte goedkope opslag kunnen bieden. Voor privégebruik wordt die opslag in "the cloud" tot een bepaalde limiet vaak zelfs gratis aangeboden.

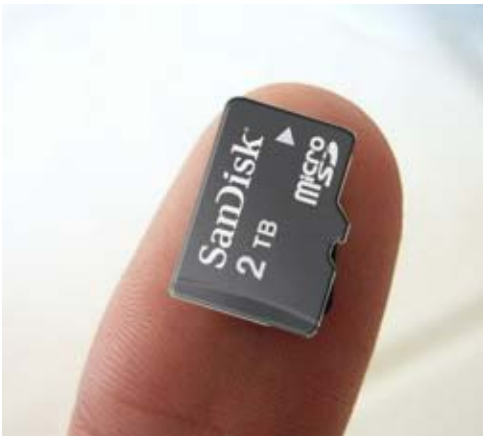

Fig. 2.8 De omvang van een 5 MB harde schijf van IBM anno 1956, vergeleken met een microSD geheugenkaartje van 2 TB anno 2013

2.3.5 Big data

De begrippen "veel" en "data" komen natuurlijk ook samen in de combinatie "big data", één van de buzz-words van de laatste paar jaar. Daarbij gaat het om zowel cijfermatige als tekstuele gegevens die uit digitale bedrijfsprocessen worden verzameld en om gegevens uit gebruikers-interactie en sociale media op internet. De grote spelers, Google, Bing, Facebook, Twitter, Amazon, beheren uiteraard ook al gigantische hoeveelheden data over ons zoek-, klik-, tweet-, like- en browsegedrag. In veel gevallen zijn die big data nog geen strikt gestructureerde gegevens, zodat data- en text-mining technieken moeten worden toegepast om hieruit betekenis en vooral trends en correlaties te kunnen afleiden. Als linked open data via internet beschikbaar gestelde datasets worden soms ook wel tot "big data" gerekend. Niet elk van die datasets hoeft echter "big" te zijn. Bovendien gaat het hier wel om gestructureerde gegevens, waaraan zelfs al op gestandaardiseerde wijze betekenis is toegekend. In elk geval zullen de totale hoeveelheden aan big data de komende jaren alleen nog maar "bigger" worden, vooral ook omdat al onze apparaten in het "internet of things" daar ook nog eens klokrond gigantische hoeveelheden data aan toe gaan voegen.

2.4 Trend 4: Meer transparantie

Een vermeldenswaardige trend van de laatste jaren ten aanzien van wetenschappelijke informatie kan gevangen worden onder de term "transparantie". Daarbij gaat het enerzijds om reguliere wetenschappelijke publicaties, waarvoor een tendens bestaat om die voor iedereen vrij beschikbaar te willen maken - de "open access" beweging. Anderzijds is er een steeds sterkere roep de onderzoeksdata die aan die publicaties ten grondslag liggen, zorgvuldig op te slaan en uiteindelijk ook beschikbaar te stellen. Beide tendensen zorgen dat wetenschappelijke resultaten in principe voor een brede groep belangstellenden toegankelijk worden - in principe voor elke spreekwoordelijke "belastingbetaler". Nu "Creative Commons" op allerlei terreinen gemeengoed wordt, is het niet meer dan logisch dat dat ook in de wetenschap gebeurt.

2.4.1 Open Access

Open Access is lang een beweging geweest waar vooral bibliothecarissen voor warm liepen. Zij waren het immers die het meest direct geconfronteerd werden met de steeds hogere tarieven die commerciële uitgevers vroegen voor abonnementen op hun wetenschappelijke tijdschriften. Zij benadrukten hoe universiteiten en onderzoeks-instituten in feite viermaal betalen voor door hen zelf geleverde inspanningen. Zij financieren het onderzoek waarop de publicaties gebaseerd zijn. Zij betalen het salaris van de onderzoekers terwijl die hun artikelen schrijven. Zij leveren de peer reviewers die de publicaties (van andere auteurs) beoordelen en becommentariëren, zonder daar door de uitgevers voor betaald te krijgen. En vervolgens moeten zij ook nog eens hoge abonnementsgelden betalen om toegang te krijgen tot hun eigen artikelen (en dan uiteraard ook die van andere onderzoekers).
Intussen begint ook bij onderzoekers zelf de weerstand tegen dit model toe te nemen. Probleem is echter dat nogal wat Open Access tijdschriften nu hoge bedragen van de auteurs vragen om hun artikelen gepubliceerd te krijgen. Bovendien is een aantal zogenaamde "predatory publishers" ontstaan, die net als de gevestigde commerciële uitgevers op het geld van de onderzoekers uit zijn, maar daar geen kwaliteit en nauwelijks toegevoegde waarde tegenover stellen - slecht uitgegeven tijdschriften met een lage impactfactor. Juist die zogenaamde impactfactor is op dit moment nog een hinderpaal op weg naar meer Open Access, omdat onderzoek financierende instanties de kwaliteit van onderzoek - bij gebrek aan beter - afmeten aan aantallen publicaties in "high impact" tijdschriften. Dit had lang een zichzelf in stand houdend effect, dat nu geleidelijk doorbroken lijkt te worden door introductie van alternatieve methoden om impact te bepalen, de zogenaamde altmetrics, waarbij bijvoorbeeld ook weerslag in sociale media kan worden betrokken. Ook zijn er aanwijzingen dat publicaties die vrij beschikbaar zijn - dus ook bijvoorbeeld in ontwikkelingslanden gelezen kunnen worden - vaker geciteerd worden en dus vanzelf een hogere impact krijgen.
Bij de andere vorm van Open Access, wordt wel in gewone tijdschriften gepubliceerd, maar worden digitale kopieën van die publicaties - eventueel na een door de uitgever vastgestelde embargo-periode - in institutionele repositories van de betreffende universiteiten beschikbaar gesteld. Ook dat zorgt dat publicaties - onder andere via Google Scholar - makkelijk toegankelijk zijn voor wie niet onder universitaire abonnementen valt.
De trend van Open Access beperkt zich overigens niet tot alleen tijdschriftartikelen. Hij breidt zich intussen ook aarzelend uit naar de wereld van het wetenschappelijke boek.

2.4.2 Toegankelijkheid van onderzoeksdata

Ook als de uitkomsten van onderzoek - in welke discipline dan ook - zijn verwerkt, geanalyseerd, geïnterpreteerd en uiteindelijk in publicaties zijn gerapporteerd, blijven de onderliggende onderzoeksdata van belang. Beschikbaarheid daarvan is vooral van nut voor:

– verantwoording en controle van onderzoek en daarop gebaseerde analyses en conclusies,
– hergebruik van data ten behoeve van nieuwe analyses of combinatie met nieuwe of andere data.

Zo bieden uitgevers van wetenschappelijke artikelen vaak al mogelijkheden om naast een artikel ook de betreffende datasets digitaal te publiceren of er op zijn minst vanuit het betreffende artikel naar te linken. Vanuit een algemenere invalshoek staan beheer en archivering van onderzoeksdata ook al enkele jaren sterk in de belangstelling - vooral in de Angelsaksische landen (Sieverts 2011). Dat heeft onder meer geresulteerd in het "Curation Lifecycle Model" van het Britse Data Curation Center (DCC).

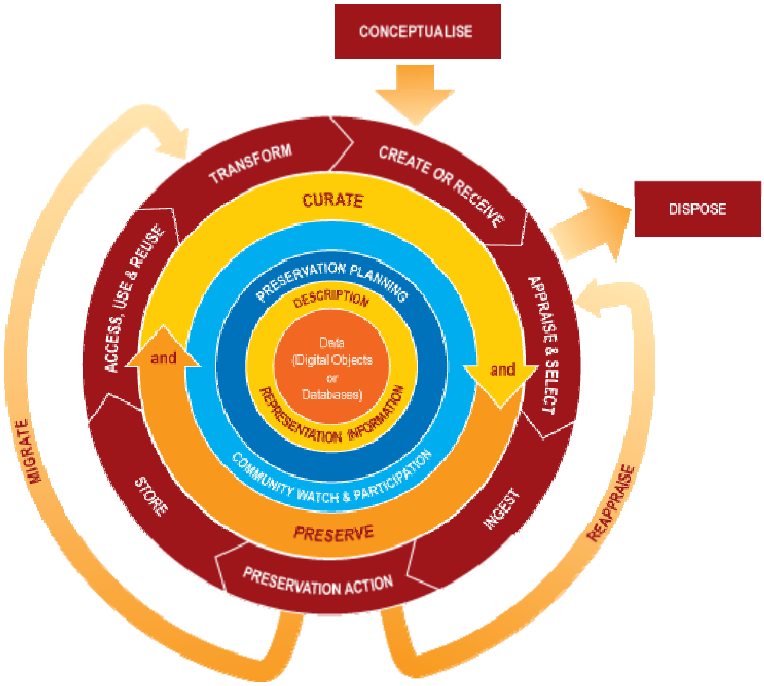

Fig. 2.9 Schematische voorstelling van het Curation Lifecycle Model (http://www.dcc.ac.uk/resources/curation-lifecycle-model).

In de loop van de levenscyclus van onderzoeksdata veranderen de rol, het gebruik en het beheer van onderzoeksdata geleidelijk. Men spreekt daarom wel van een curation continuum. Voor een modelmatige beschrijving worden "curation boundaries" aangegeven, waar data een andere fase ingaan (zie Fig. 2.10). In een eerste fase zijn gegevens nog helemaal in het privé-domein van de onderzoekers: ze beheren hun data zelf, ze werken er nog aan, vullen ze aan, analyseren ze, enzovoort. In een tweede fase zijn de data beschikbaar voor samenwerking met partners binnen of buiten de organisatie. Ze moeten dan dus zo worden opgeslagen, dat die partners er ook toegang toe hebben, en bovendien van metadata voorzien, zodat die hun betekenis kunnen interpreteren. Wat betekenen de getallen, waaraan is gemeten, wat voor proefpersonen

zijn gebruikt, wat wordt op de filmpjes waargenomen? Voor buitenstaanders mogen ze dan meestal nog niet toegankelijk zijn. Hoe streng men daarbij wil zijn, kan ook te maken hebben met eventuele commerciële en concurrentiebelangen. In een derde fase kunnen de data in het publieke domein terechtkomen: als er geen redenen zijn die dat ongewenst maken, kan iedereen er gebruik van maken. Ze moeten dan duurzaam bewaard worden en ook moet er met persistente identifiers naar gelinkt kunnen worden. Bij elke overgang naar een volgende fase zullen ook beslissingen genomen moeten worden ten aanzien van toegankelijkheid, geschiktheid van metadata en zelfs van de noodzaak tot verder bewaren. Niet alle ooit geproduceerde data hoeven in dat derde stadium terecht te komen.

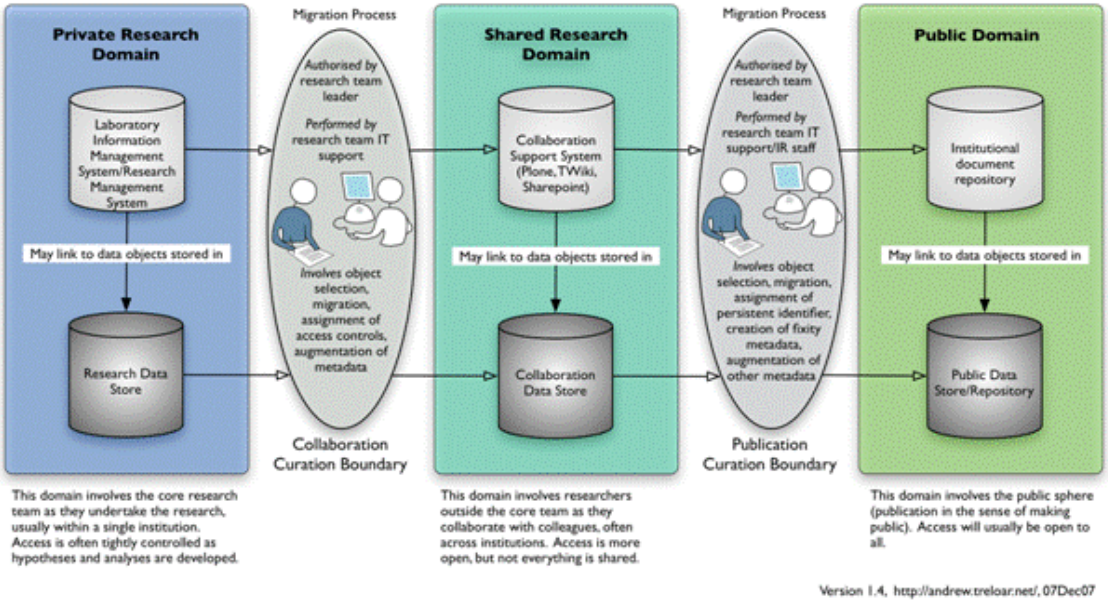

Fig. 2.10 De drie fases van het beheer van onderzoeksgegevens en de "curation boundaries" daartussen (http://ands.org.au/guides/curation.continuum.pdf).

In Nederland is de belangstelling hiervoor wat later op gang gekomen. Wel was DANS (Data Archiving and Networked Services, vallend onder KNAW en NWO) al lang actief om vooral de resultaten van sociaal wetenschappelijk onderzoek, zoals enquêteresultaten te beheren. De Universiteitsbibliotheek Utrecht was in 2009 een proefproject begonnen om onderzoekers hun eigen datasets te laten archiveren. SURF had sinds ongeveer dezelfde tijd een Onderzoeksdata Forum, intussen omgevormd tot de Special Interest Group Research Data. Toch heeft pas de commotie rond het grootscheepse bedrog door Diederik Stapel er voor gezorgd dat dit onderwerp echt in de belangstelling is komen te staan en dat - zelfs internationaal - gedacht wordt over een verplichting tot het - waar nodig onder voorwaarden van privacy of vertrouwelijkheid - beschikbaar stellen van onderzoeksdata.
Dit heeft het instellen van depots voor onderzoeksdata en discussies over het noodzakelijke metadateren van datasets in een stroomversnelling gebracht. Hoewel de vraagstukken die hierbij spelen per discipline heel sterk verschillen, zijn al steeds grotere hoeveelheden onderzoeksdata beschikbaar. De vindbaarheid van die datasets blijft daar op dit moment nog sterk bij achter. Er is nog geen Google Scholar voor datasets.

3 De informatiewereld in 2023

Bij de keuze van de in de vorige paragraaf beschreven trends ben ik in de eerste plaats uitgegaan van mijn eigen interessegebied. Daaruit heb ik bovendien die aspecten gekozen, waarvan ik verwacht dat ze ook de komende jaren zullen doorzetten. Wanneer we tien jaar terugkijken en de situatie van toen vergelijken met die van nu, dan zien we grote, maar niet werkelijk revolutionaire veranderingen. Het lijkt daarom geen gewaagde veronderstelling dat ook de komende tien jaar vooral een verdere evolutie zal plaats vinden en dat niet ineens een volstrekt ander informatielandschap zal ontstaan, waarin de huidige trends niet meer doorwerken. Dat betekent overigens niet dat hier geen verdere overdenkingen over 2023 meer zullen volgen. Die zullen echter maar voor een deel betrekking hebben op de hier eerder beschreven trends.

3.1 Cyberbrain

Recent stond op ReadWrite blog een bijdrage onder de titel "What if Google Could Think Like You Do?" (Hachman 2013). Die begint met de zin "The next *space race* might be the race to develop a synthetic model of the human brain - one that Google and Microsoft will participate in". Modelleren (en begrijpen) hoe de hersenen werken is iets waarmee onderzoekers op het gebied van supercomputing en kunstmatige intelligentie zich al langer bezig hielden. Nieuw was dat een zoekmachinegigant daar nu ook al mee in verband werd gebracht. Hachman verwijst in zijn bijdrage naar een DARPA-initiatief onder de naam SYNAPSE, waarin een nieuw soort "neuromorphic" computerarchitectuur wordt ontwikkeld (Synapse 2012). Met het complexer worden van de uit te voeren taken zou de complexiteit van die machine veel minder snel hoeven toe te nemen dan bij meer klassieke (von Neumann-) computers (zie Fig. 3.1). Zo zou men uiteindelijk een Cyberbrain moeten kunnen ontwikkelen.

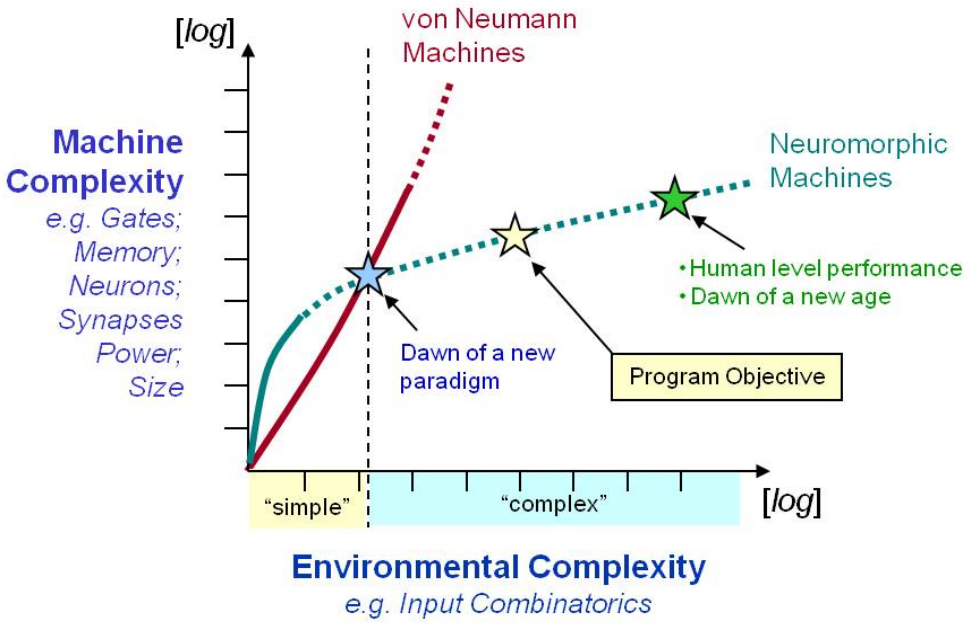

Fig. 3.1 Relatie tussen de complexiteit van een voor een Cyberbrain te ontwikkelen machine en de complexiteit van de daarmee op te lossen problemen.

Hoewel de horizontale as in Fig. 3.1 geen tijdschaal is, is hij daar natuurlijk wel aan gerelateerd. Zo moet het "Program Objective", het doel van het huidige onderzoek-programma al rond 2019 gehaald worden. Hoe ver in de toekomst men realisatie van de uiteindelijk nagestreefde "Human level performance" inschat, viel niet makkelijk te achterhalen. Wel wordt rond 2023 een "full human brain simulation" verwacht - als er althans nog voldoende subsidie komt.
Google blijkt niet rechtstreeks bij dit project betrokken, maar heeft wel recent futuroloog Ray Kurzweil als onderzoeksdirecteur aangesteld. Die moet meewerken aan de eigen ontwikkeling van een kunstmatig brein (Levy 2013). Kurzweil heeft al eerder ideeën geventileerd over zo'n brein, dat ook wel wordt aangeduid met de term "singularity". Aan de haalbaarheid van die ideeën, wordt door neurowetenschappers overigens weinig geloof gehecht (Regalado 2013). Niettemin bieden deze ontwikkelingen interessante perspectieven dat machines tegen 2023 in elk geval nog aanzienlijk beter kunnen begrijpen wat we zoeken en waarin we zoeken.

3.2 Six things in 2023

Op de ReadWrite blog werd - in het kader van het tienjarig bestaan - bewust tien jaar vooruitgekeken naar 2023. Omdat dit een weblog is waar de afgelopen tien jaar veel algemene ontwikkelingen rond internet aan de orde kwamen en (vaak juist) voorspeld werden, is het goed ook eens te kijken naar de "Six Things ReadWrite Will Cover In 2023" (Thomas 2013). Dat zijn:

1. ***Anticipatory systems***. Computers wachten niet meer op input, zoals zoekopdrachten, maar anticiperen zelf - op basis van eerdere ervaring - wat hun gebruiker zal willen. Overigens geen heel revolutionaire voorspelling, want Google-Now en Apple's Siri doen dit in feite nu al.
2. ***No gadgets anymore***. We hebben geen smartphones, tablets of slimme horloges meer nodig. We zullen overal (flexibele) beeldschermen ter beschikking hebben, waarop alle informatie in het klein of in het groot geprojecteerd wordt.
3. ***Ambient electricity***. Processoren worden vooral energiezuiniger in plaats van nog sneller. Daardoor wordt het goed mogelijk om de daarvoor benodigde batterijen via radiogolven op te laden.
4. ***Self driving cars***. Voertuigen zullen geen bestuurdersplaats en stuurwiel meer hebben omdat ze automatisch achter elkaar gekoppeld worden, tot treintjes die naar de juiste bestemming geleid worden.
5. ***Everybody a programmer***. Professioneel programmeren zal geen betaald beroep meer zijn, omdat iedereen het kan, doordat het een drag-and-drop proces wordt dat niet ingewikkelder is dan het downloaden van een App.
6. ***Data cooled undersea***. Doordat de groei van de dataproductie onverminderd aanhoudt, moeten de "cloud" datacentra steeds goedkoper van elektriciteit en koeling worden voorzien. In plaats van naast het water komen ze onder water.

Hoewel niet alle zes deze onderwerpen rechtstreeks aan informatievoorziening en informatielandschap gerelateerd zijn, bieden ze een interessant algemeen perspectief op de komende informatiemaatschappij.

3.3 Google Glass

De Googlebril - Google Glass - lijkt op dit moment in de eerste plaats een hebbedingetje, vooral omdat hij in Nederland op het moment van schrijven nog niet verkrijgbaar is. In Amerikaanse blogs kom je intussen enthousiaste berichten tegen,

zoals "This is why Google Glass is the future" (Ulanoff 2013). Anderzijds is er ook wel kritiek. Een korte samenvatting:

- Wie gaat er nu met zo'n raar uitziend brilletje rondlopen?
- De spraakaansturing werkt nog niet goed.
- Het is een niche-product dat maar voor een kleine doelgroep interessant is.
- Het is gevaarlijk voor gebruikers, omdat het nog meer afleidt dan een smartphone.
- Mensen willen nog niet in zo'n semi-virtuele wereld leven.
- Ze zijn veel te duur.

Degene die deze bezwaren verzameld had, verwacht dat dit soort apparaten pas tegen 2023 "mainstream" wordt (Hanson 2013). Tegen die tijd zullen inderdaad wel wat van de problemen van techniek en sociale acceptatie zijn opgelost. En ook moeten tegen die tijd wel wat interessante toepassingen tot ontwikkeling zijn gekomen, waar we in de informatiemaatschappij van dat moment wat aan hebben. Bibliotheektoepassingen die op dit moment de revue passeren liggen nog merendeels in het verlengde van nu al bestaande diensten en mogelijkheden, al lijken ze met Google Glass wel een stuk gebruiksvriendelijker te realiseren (Kroski 2013). Hoe dan ook zal gepersonaliseerde, context- en locatiegevoelige, anticiperende informatievoorziening vrij zeker tot de succesvolle toepassingen blijven behoren.

3.4 En verder nog

Hoewel Google al bijna 15 jaar bestaat en Facebook ook al ruim acht jaar, is er geen garantie dat die huidige grootmachten over tien jaar nog even toonaangevend zullen zijn. Diensten als Facebook en Twitter lijken wat trendgevoeliger te zijn dan een basisfunctionaliteit als "zoeken" (ook al is het totale zoekvolume recent wat gedaald). Bij dat zoeken - al dan niet gedomineerd door Google - zal semantiek in 2023 een nog belangrijker rol spelen dan het nu al doet. Die rol zal er vooral op drie terreinen zijn:

- het - al dan niet geautomatiseerd - vooraf meegeven van betekenis aan op internet gepubliceerde informatie, zoals dat nu al in toenemende mate met "embedded metadata" gebeurt;
- het beschikbaar zijn van grote hoeveelheden gestructureerde, gevalideerde, feitelijke informatie, zoals dat nu al van de grond komt met Knowledge Graph, linked data, DBpedia, Wikidata etc.;
- betere "on-the-fly" analysetechnieken voor het afleiden van betekenis in niet al vooraf semantisch gekarakteriseerde informatie.

Er zal zeker een verdere perfectionering hebben plaats gevonden van het analyseren van ons sociale mediagedrag, in combinatie met op andere wijze gegenereerde big data die op personen te herleiden zijn, bijvoorbeeld data uit allerlei autonoom op internet aangesloten apparatuur in ons huis en op ons lichaam. Tegelijkertijd zal dat echter een zware wissel trekken op regelgeving rond de bescherming van privacy en persoonlijke levenssfeer, en vooral ook op handhaving van die regels.
Een onderwerp dat hier nog niet expliciet aan de orde is geweest, is het zoeken binnen (de informatie van) organisaties, "enterprise search". Voor een deel zijn daarvoor dezelfde ontwikkelingen te voorzien als voor websearch. Maar er is ook verschil. Enerzijds is er verschil in schaalgrootte met "het hele internet"; anderzijds worden aan het zoeken vaak andere eisen gesteld - meestal mogen geen documenten gemist worden en dat ene document moet beslist boven tafel komen. Daarvoor zijn betere "brute force" computertechnieken misschien al voldoende, al zullen bij enterprise search in 2023 ook semantiek en sociale technieken hun intrede hebben gedaan.

4 Consequenties en maatregelen

In mijn bijdrage zal ik tot slot nog heel beperkt ingaan op de verdere consequenties die de hier gegeven toekomstvisie met zich mee brengt en op de maatregelen die dat wellicht nodig maakt.

Voor wie in de informatiemaatschappij van 2023 werkzaam is, in welke functie dan ook, zal "Information Literacy" - zelfs nog meer dan nu al - een kerncompetentie zijn. Maar die zal aangevuld moeten zijn met een voldoende niveau van "Technological Literacy". Dat is niet bedoeld op het niveau van bits en bytes of programmeertalen, aangezien het gebruiksgemak van systemen ook de komende tien jaar nog verder zal toenemen. Maar wel op een meer algemeen conceptueel niveau, dat het mogelijk maakt de persoonlijke informatieomgeving in hoge mate zelf vorm te geven. Deze competenties zullen dus centraal onderdeel van ieders scholing moeten zijn.

Semantiek - maar dan niet meer in de vorm van "onderwerpsontsluiting" - zal nog meer een rol spelen dan het nu al begint te doen. Iedereen zal dus een metadataris dienen te zijn, voor wie gebruik en toepassing van metadata - in welke vorm dan ook - geen geheimen heeft. Overigens is dat tegen die tijd een gewoon onderdeel van de al eerder genoemde information literacy.

In het licht van de nog steeds voortdurende exponentiële groei van onze informatie- en data-productie, zullen we toch al moeten anticiperen op een toekomst waarin er noodgedwongen "Grenzen aan de groei" zullen komen. Dat zal onder meer betekenen dat we niet alles meer moeten willen opslaan en bewaren. Er zal geselecteerd moeten worden en we zullen gecontroleerd moeten weggooien. Voor archivarissen klinkt dat vertrouwd, maar het zal ook nodig zijn op terreinen die nu nog niet tot het domein van de archivaris behoren.

5 Bronnen

Brouwers, Lucas (2011). Meer en meer wetenschap - NRC Handelsblad, 12 maart 2011

Connaway, Lynn Silipigni; Dickey, Timothy J. and Radford, Marie L. (2011). “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, Vol. 33 (2011) 179-190 (https://www.oclc.org/content/dam/research/publications/library/2011/connaway-lisr.pdf , geraadpleegd 24 april 2013)

Cyganiak, Richard (2011) - The Linking Open Data cloud diagram (http://richard.cyganiak.de/2007/10/lod/ , geraadpleegd 24 april 2013)

Das, Abhishek and Jain, Ankit (2012). Indexing The World Wide Web: The Journey So Far - In: Next Generation Search Engines: Advanced Models for Information Retrieval, IGI-Global (http://research.google.com/pubs/archive/37043.pdf , geraadpleegd: 8 maart 2013)

De Solla Price, Derek (1963). Little science, big science. New York: Columbia University Press. ISBN 0-231-08562-1.

Gallagher, Sean (2012). How Google and Microsoft taught search to “understand” the Web. In: ArsTechnica Technology Lab. 7 juni 2012 (http://arstechnica.com/information-technology/2012/06/inside-the-architecture-of-googles-knowledge-graph-and-microsofts-satori/ , geraadpleegd 25 april 2013)

Hachman, Mark (2013). What if Google Could Think Like You Do? - In: ReadWrite Enterprise. 19 february 2013 (http://readwrite.com/2013/02/19/what-if-google-could-think-like-you-do , geraadpleegd 25 april 2013)

Hanson, Arik (2013). Google Glass: Why it won’t go mainstream until 2023. In: Communications Conversations. 17 maart 2013 (http://www.arikhanson.com/2013/03/27/google-glass-why-it-wont-go-mainstream-until-2023/ , geraadpleegd 26 april 2013)

Havard (2013). Historical Cost of Computer Memory and Storage. (http://hblok.net/blog/posts/2013/02/13/historical-cost-of-computer-memory-and-storage/ , geraadpleegd: 24 april 2013)

Huurnink, Bouke; Snoek, Cees G. M.; de Rijke, Maarten and Smeulders, Arnold W. M. (2012). Content-Based Analysis Improves Audiovisual Archive Retrieval. IEEE Transactions on Multimedia, Vol. 14, No. 4 (2012) 1166-1178 (http://staff.science.uva.nl/~cgmsnoek/pub/huurnink-archive-tmm.pdf , geraadpleegd 24 april 2013)

Landry, Tommy (2013). Semantic Web: Are You Taking Advantage of Semantic Search? - In: Search Engine Journal (April 20, 2013)

(http://www.searchenginejournal.com/semantic-web-are-you-taking-advantage-of-semantic-search/62047/ , geraadpleegd 25 april 2013)

Kalakota, Ravi (2011). Big Data Infographic and Gartner 2012 Top 10 Strategic Tech Trends - In: Business Analytics 3.0. 11 november 2011 (http://practicalanalytics.wordpress.com/2011/11/11/big-data-infographic-and-gartner-2012-top-10-strategic-tech-trends/ , geraadpleegd 8 maart 2013)

Kroski, Elyssa (2013). 7 Things Libraries Can Do with Google Glass. In: OEDb iLibrarian. 18 april 2013 (http://oedb.org/blogs/ilibrarian/2013/6-things-libraries-can-do-with-google-glass/ , geraadpleegd 27 april 2013)

Levy, Steven (2013). How Ray Kurzweil Will Help Google Make the Ultimate AI Brain. In: Wired Business. 25 april 2013 (http://www.wired.com/business/2013/04/kurzweil-google-ai/ , geraadpleegd 26 april 2013)

Library linked data incubator group wiki (2012) (http://www.w3.org/2005/Incubator/lld/wiki/Main_Page , geraadpleegd 24 april 2013)

Lyman, Peter and Varian, Hal R. (2000). How much information? (http://www2.sims.berkeley.edu/research/projects/how-much-info/ , geraadpleegd: 8 maart 2013)

Lyman, Peter and Varian, Hal R. (2003). How much information? 2003 (http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf , geraadpleegd: 8 maart 2013)

Pariser, Eli (2011). The Filter Bubble: What the Internet Is Hiding from You. Penguin Press (New York, May 2011) ISBN 978-1-59420-300-8

Regalado, Antonio (2013). The Brain Is Not Computable. In: MIT Technology Review. 18 february 2013 (http://www.technologyreview.com/view/511421/the-brain-is-not-computable/ , geraadpleegd: 8 maart 2013)

Sengupta, Somini (2013). Facebook Shows Off New Home Page Design, Including Bigger Pictures - New York Times Technology. 7 maart 2013 (http://www.nytimes.com/2013/03/08/technology/facebook-shows-off-redesign.html , geraadpleegd: 8 maart 2013)

Sieverts, Eric (2011). De cirkel van onderzoeksdata. In: Toegang tot onderzoeksdata. Stichting SURF. 2011. p. 6-11. (http://www.surf.nl/nl/themas/openonderzoek/permanentetoegangtotdata/Documents/ToegangOZdataweb.pdf , geraadpleegd 26 april 2013)

Sieverts, Eric (2012). 40 jaar informatiegebruik; informatie vinden en selecteren in tijden van informatieovervloed - Afscheidscollege, 13 januari 2012 (Amsterdam) (http://www.slideshare.net/sieeg/40-jaar-informatiegebruik , geraadpleegd: 8 maart 2013)

Snoek, Cees G. M. and Worring, Marcel (2008). Concept-Based Video Retrieval. In: Foundations and Trends in Information Retrieval. Vol. 2, No. 4 (2008) 215-322, (http://staff.science.uva.nl/~cgmsnoek/pub/snoek-concept-based-video-retrieval-fntir.pdf , geraadpleegd 24 april 2013)

Starr, Barbara (2012). How Search & Social Engines Are Using Semantic Search - In: SearchEngineLand. 26 september 2012 (http://searchengineland.com/semantic-search-what-is-it-how-are-major-search-and-social-engines-use-it-part-1-133160 , geraadpleegd 24 april 2013)

Starr, Barbara (2013). Semantic & Graph-Based Search: The Future Face Of Search - In: SearchEngineLand. 25 april 2013 (http://searchengineland.com/semantic-graph-based-search-the-future-face-of-search-156461 , geraadpleegd 25 april 2013)

Synapse (2012). Systems of Neuromorphic Adaptive Plastic Scalable Electronics (http://www.darpa.mil/Our_Work/DSO/Programs/Systems_of_Neuromorphic_Adaptive_Plastic_Scalable_Electronics_(SYNAPSE).aspx , geraadpleegd: 8 maart 2013)

Thomas, Owen (2013). Six Things ReadWrite Will Cover In 2023. In: ReadWrite Ten. 19 april 2013 (http://readwrite.com/2013/04/19/rw10-readwrite-2023 , geraadpleegd 26 april 2013)

Ulanoff, Lance (2013). This is why Google Glass is the future. In: Mashable. 30 april 2013 (http://mashable.com/2013/04/30/google-glass-future/ , geraadpleegd 30 april 2013)

Youtube (2012). Statistics (http://www.youtube.com/yt/press/statistics.html , geraadpleegd: 8 maart 2013)