Conferentie informatiewetenschap 2000, de Doelen, Utrecht, 5 april 2000

Eindhoven University of Technology
Department of Mathematics and Computing Science

Conferentie Informatiewetenschap 2000

De Doelen Utrecht
5 april 2000

Proceedings

Edited by P. van der Vet and P. De Bra

00/20

Editors: prof.dr. J.C.M. Baeten
Prof.dr. P.A.J. Hilbers

Reports are available at:
http://www.win.tue.nl/win/cs

Computing Science Reports 00/20
Eindhoven, December 2000

Conferentie Informatiewetenschap 2000

De Doelen Utrecht
5 april 2000

Proceedings

Edited by P. van der Vet and P. De Bra

Digitalisering en digitale archivering van de Nederlandse volkstellingen, 1795-1971

Peter Doorn
Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI)
Joan Muyskenweg 25
Postbus 95110
1090 HC Amsterdam
T +31 20 4628 606
F +31 20 6658 013
E peter.doorn@niwi.knaw.nl

Inhoud

1. Inleiding

Het Centraal Bureau voor de Statistiek (CBS) en het Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI) werken sinds 1997 samen aan het project 'Digitalisering Nederlandse Volkstellingen 1795-1971'. In 1795 werd de eerste landelijke volkstelling in ons land gehouden en in 1971 de laatste. De volkstellingen (VT's) bevatten een schat aan historische, sociaal-economische, demografische en culturele gegevens.
In totaal zijn in de laatste twee eeuwen een kleine 200 banden met ruim 42.000 bladzijden aan tabellen (en toelichtingen daarop) gepubliceerd. In de bibliotheek en het archief van het CBS bevinden zich daarnaast nog enkele honderdduizenden bladen met ongepubliceerd materiaal over de laatste drie Volkstellingen (1947, 1960 en 1971). Van de tellingen van 1960 en 1971 zijn digitale bestanden bewaard gebleven.
De digitalisering van de Nederlandse volkstellingen is gesubsidieerd door het fonds Innovatie Wetenschappelijke Informatievoorziening (IWI) van SURF en door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Gebiedsbestuur voor de Maatschappij- en gedragswetenschappen en Wetenschappelijk Statistisch Agentschap).

Het belang van (digitale) volkstellingen

Nationale volkstellingen behoren tot de meest elementaire informatiebronnen over de toestand in een land. Naast de omvang van de bevolkingsgrootte bevat de volkstelling doorgaans informatie over de structurele kenmerken van een land, zoals leeftijd, geslacht, burgerlijke staat, levensbeschouwing, huishoudenssituatie, beroepswerkzaamheid en nationaliteit. In diverse jaren is de volkstelling gecombineerd gehouden met een beroepstelling en een woningtelling.
De eerste algemene volkstelling in Nederland vond plaats in 1795 onder de Bataafse Republiek. Sedert 1829/1830 is er sprake van periodieke tellingen, die eens in de 10 jaar werden gehouden. De telling van 1940 werd in verband met de oorlog uitgesteld tot 1947. Na de veertiende telling in 1971 is in Nederland geen volkstelling meer gehouden als gevolg van de toegenomen privacy-bewustheid (weigering tot medewerking) van de bevolking.
Van de Nederlandse volkstellingen 1795-1971 is slechts een beperkt aantal exemplaren in Nederland aanwezig. Afgezien van de bibliotheek van het CBS zelf, beschikken de meeste universiteitsbibliotheken over een min of meer complete statistische collectie. De volkstellingen hebben altijd een belangrijke rol gespeeld in het historisch en sociaal-wetenschappelijk onderzoek. Veel van de gepubliceerde volkstellingsboeken verkeren inmiddels in slechte staat. Nu de volkstellingen digitaal beschikbaar zijn, kunnen de originele boeken in de kast blijven staan en hoeven ze niet verder te lijden van het gebruik.
De digitale volkstellingen vormen een belangrijk instrument voor historisch en sociaal-wetenschappelijk onderzoek. In tal van landen worden of zijn historische censusprojecten uitgevoerd. In diverse gevallen zijn deze projecten gebaseerd op het oorspronkelijke basismateriaal, waardoor databases op individueel niveau zijn opgebouwd. In Nederland is het basismateriaal van de volkstellingen niet (centraal) bewaard gebleven. Er zijn wel gedeelten van het basismateriaal op individueel niveau bewaard gebleven bij (sommige) gemeentelijke archiefinstellingen, aangezien de gemeenten met de uitvoering van de volkstellingen waren belast. De gepubliceerde tellingen bieden vaak zeer gedetailleerde informatie. De Volkstelling 1899 (inclusief beroepstelling en woningtelling) omvat ca. 10.000 pagina's met tabellen. Eén tabelpagina bevat gemiddeld bijna 700 cellen. De gepubliceerde Volkstelling 1899 telt in totaal dus zo'n zeven miljoen gegevens. Bijvoorbeeld, van iedere gemeente worden per buurt of wijk en type woonruimte gegevens gepresenteerd, die vaak tot op het individu herleidbaar zijn. Om uiteenlopende redenen is ervoor gekozen om de tellingen van 1899 als eerste integraal te digitaliseren. Dit wil zeggen dat niet alleen digitale opnamen (*images*) zijn gemaakt van alle bladzijden, maar dat alle gegevens zijn opgenomen in databases, die volledige doorzoekbaar en raadpleegbaar zijn.

- De Volkstelling van 1899 is één van de meest uitvoerige tellingen. Zij staat bekend als een kwalitatief zeer goede telling. Er was een combinatie met een beroepstelling, met een zeer uitvoerige beroepenclassificatie, die het uitgangspunt heeft gevormd voor indelingen in de 20e eeuw. Er doen zich wel aanzienlijke verschillen voor met de beroepenclassificatie van 1889, die in historisch onderzoek veelal wordt gebruikt voor 19e eeuwse en eerdere sociaal-economische stratificaties.
- 1899 ligt vrijwel halverwege de periode 1795-heden. De telling is enerzijds 'modern' van opzet, maar anderzijds vergelijkbaar met de eerdere tellingen in de 19e eeuw. De ervaringen van het project zullen dan ook goed bruikbaar zijn voor de inhoudsconversie van de overige volkstellingen in een later stadium.
- Het CBS is opgericht in 1899. Het was de eerste telling die onder verantwoordelijkheid van het toen kersverse bureau werd uitgevoerd. Het eeuwfeest van het CBS vormt een uitstekende gelegenheid voor een heruitgave van de eerste CBS-telling, maar nu in digitale vorm.

De digitale volkstellingen vormen niet alleen een uiterst bruikbaar instrument voor wetenschappelijk onderzoek, maar zullen tevens tot de verbeelding van het publiek in ruimere zin spreken. Hierbij wordt gedacht aan amateur-historici, overheden, studenten en scholieren.

Historische digitale volkstellingen in het buitenland

In tal van andere landen worden eveneens projecten uitgevoerd, waarin historisch volkstellingsmateriaal digitaal wordt gemaakt. Dit overzicht biedt een selectie van ingangen op digitaal beschikbare volkstellingen in enkele andere landen. Het pretendeert niet compleet te zijn, en de geboden Internet-adressen kunnen zijn verouderd.
In Scandinavië is het materiaal van bepaalde tellingen integraal raadpleegbaar op individueel niveau, bijvoorbeeld de census van Noorwegen in 1801. In Denemarken wordt met vrijwilligers gewerkt aan de totale invoer van alle volkstellingen tussen 1787 en 1916. Grote hoeveelheden gegevens zijn reeds beschikbaar bij het Deense Data Archief in Odense.
In de Verenigde Staten zijn belangrijke censusgegevens tussen 1790 en 1970 beschikbaar in het data-archief van het ICPSR te Ann Arbor op geaggregeerd niveau. Bij de Universiteit van Virginia is een gebruikersvriendelijke toegang op deze gegevens gemaakt. In het IPUMS-project (Integrated Public Use Microdata Series) van de Universiteit van Minnesota worden steekproeven van de Amerikaanse volkstellingen tussen 1850 en 1990 toegankelijk gemaakt. De oorspronkelijke digitale bestanden sinds 1970 van het Amerikaanse Bureau of the Census zijn gearchiveerd in het Data Access and Dissemination System (DADS) project.
Voor tal van Amerikaanse regio's en deelstaten zijn projecten gaande waarin (delen van) historische volkstellingen worden ontsloten. Er bestaan ook veel door genealogen en amateur-historici opgebouwde bestanden met census-gegevens. In het project 'Census Online' worden koppelingen naar meer dan 2000 Web-sites in de Verenigde Staten geboden.
In Ierland is een project uitgevoerd om de belangrijkste gegevens uit de gedrukte volkstellingen tussen 1821 en 1971 in een database op te nemen. De Database of Irish Historical Statistics claimt nu ongeveer 150 miljoen gegevens te bevatten.
In de Historical Data Service, een onderdeel van de Engelse Arts and Humanities Data Service en het Data Archief, wordt een rijke collectie historische statistieken en volkstellingen toegankelijk gemaakt in de 'Great Britain Historical Database'. Gewerkt wordt aan de ontsluiting van de Volkstelling van 1851 op individueel niveau. In het UK Data Archive is de 1881 Census for England, met 26 mln records, opgenomen. Deze is gedigitaliseerd door de Mormonen.
Zoals gezegd zijn er nog veel meer projecten waarin historisch-statistisch materiaal grootschalig wordt gedigitaliseerd en toegankelijk gemaakt. Het Norwegian Historical Data Centre biedt een bruikbaar overzicht van 'historical microdata around the world' (URL: http://www.isv.uit.no/seksjon/rhd/nhdc/micro.htm).

Tabel 3: Enkele historische censusprojecten in het buitenland

Land	Jaren van tellingen	Project en URL	Instituut, plaats
Noorwegen	1801	Census of Norway 1801 http://www.uib.no/hi/1801page.html	Department of History, University of Bergen
Denemarken	1787-1916	Danish Demographic Database http://ddd.sa.dk/ddd_en.htm	Danish Data Archive, Odense
VS	1790-1970	US Historical Census data browser http://fisher.lib.virginia.edu/census/	ICPSR, Ann Arbor enUniversity of Virginia Library
VS	1850-1990	IPUMS http://www.ipums.umn.edu/	University of Minnesota
VS	Links naar > 2000 Websites	Census Online http://www.census-online.com/index.html	
VS	sinds 1970	DADS project http://merrill.wwh.net/mdocs/census.html	Lawrence Berkeley National Laboratory
Ierland	1821-1971	Database of Irish Historical Statistics http://www.qub.ac.uk/ss/csr/iredb/dbhme.htm	Centre for Data Digitisation and Analysis, Queen's University of Belfast
Engeland	Diverse jaren	Great Britain Historical Database http://hds.essex.ac.uk/gbh.stm	Historical Data Service, Essex

2. Materiaalselectie

Een selectie van de te verwerken volkstellingsdelen 1795-1971 is gemaakt aan de hand van een vergelijking van de staat van de drukkwaliteit in de bibliotheken van het CBS, de Rijksuniversiteit Leiden, het Internationaal Instituut voor Sociale Geschiedenis, de Katholieke Universiteit Brabant en de Rijksuniversiteit Groningen. Behalve de delen met tabellen zijn ook belangrijke publicaties met informatie over de volkstellingen (zoals monografieën en inleidingen) verwerkt. In het algemeen bleek de drukkwaliteit van de geselecteerde boeken bevredigend. Wel zijn er delen die in alle bibliotheken dezelfde problemen vertonen (doordrukken achterzijde pagina's, krappe binding). De tellingen in 1807/1808 en 1815 zijn niet landelijk uniform en ook niet centraal beschikbaar. De kwaliteit en daarmee de bruikbaarheid voor onderzoek daarvan is wisselend. Mede op grond van externe adviezen is besloten deze tellingen niet in het project te betrekken.

Tabel 1: Aantallen delen en pagina's van alle gepubliceerde volks-, beroeps- en woningtellingen, 1795-1971 (excl.lichtdrukken)

Jaar	Pagina's	Delen
1795	191	2
1829	18	*
1840	85	1
1849	1165	12
1859	1184	3
1869	889	3
1879	2262	12
1889	10223	26
1899	9925	27
1909	4144	14
1919	191	1
1920	1953	10
1930	2353	11
1947	1325	12
1956	345	3
1960	1809	18
1971	4503	38
Totaal	**42565**	**193**

* Enkele resultaten gepubliceerd in het Statistisch Jaarboekje van 1830 en in deel 2 van de Volkstelling van 1859.

Op basis van een technisch vooronderzoek is vastgesteld dat een combinatie van microverfilming met microfilmscanning de meest efficiënte wijze van verwerking zou zijn. De volkstellingen zijn opgenomen op 35 mm film door de firma Microformat met een zo gering mogelijke verkleiningsfactor. Omdat de meeste tabellen doorlopen over twee pagina's, is er voor gekozen per *frame* twee pagina's op te nemen. De opnamen werden aangeleverd op 40 rolfilms met gemiddeld ruim 500 images per film. Afwijkende formaten (uitklaptabellen, kaarten en dergelijke) werden op een aparte film gezet.
Van een deel van de 19e eeuwse volkstellingen waren in het verleden al eens microfilms en/of fiches gemaakt, maar de kwaliteit daarvan werd als onvoldoende voor digitalisering beschouwd. Scanning op flatbed-scanners van de deels toch al kwetsbare boeken zou naast mogelijke schade ook een tragere en minder bedrijfszekere verwerking opleveren. Ook is nog overwogen om de boeken uit de band te lichten en na losbladige verwerking opnieuw in te binden, maar deze werkwijze stuitte eveneens op diverse praktische bezwaren (bijvoorbeeld het feit dat bij het loshalen van katernen de tabellen die over het pagina-midden doorlopen niet meer op elkaar aansluiten).

Oude lichtdrukken, transparanten en tapes

Naast de gepubliceerde tellingen beschikt de bibliotheek van het CBS over ca. 192.000 pagina's met uiterst gedetailleerde volkstellingsinformatie op lichtdrukken uit de periode 1947-1971. De omvang van het gelichtdrukte materiaal is dermate groot (vier- tot vijfmaal meer dan de gepubliceerde tellingen!) en de kwaliteit ervan is dermate wisselend (gedeeltelijk handgeschreven en gedeeltelijk geprint met matrix- of kettingprinter), dat de ontsluiting hiervan bijzondere aandacht vergt. Er is onderzocht op welke wijze dit materiaal geconverteerd kan worden en hoeveel kosten hiermee gemoeid zouden zijn.
De originelen, waarvan de lichtdrukken zijn gemaakt, zijn opgespoord in het archief van het CBS. Het blijkt voor de periode 1947-1960 te gaan om voorbedrukte formuleren op transparant of calqueerpapier, waarop de gegevens zijn geschreven. De staat van de originelen loopt sterk uiteen, afhankelijk van de vraag hoe vaak de originelen zijn gebruikt om er lichtdrukken van te maken. Het uiterst dunne papier is zeer bros. Scheuren zijn vaak gerepareerd met plakband.
De eerste volkstelling waarbij computerverwerking werd toegepast, was die van 1960. De bestanden zijn door het CBS enkele jaren geleden gedeponeerd bij het Steinmetzarchief (thans is dit sociaal-wetenschappelijke data-archief onderdeel van het NIWI), nadat de ponskaarten c.q. oorspronkelijke tapes opnieuw waren ingelezen op het rekencentrum SARA van de Universiteit van Amsterdam.

Bij de conversie en documentatie door het Steinmetzarchief is gebleken dat de bestanden lacunes vertonen en niet geheel overeenstemmen met de gepubliceerde volkstellingsresultaten. Het onderzoek naar de mogelijkheden van een 'digitale restauratie' van het bestand van de Volkstelling 1960 is nog gaande.
De gegevens met bijbehorende documentatie van de Volkstelling 1971 zijn bij het CBS nog aanwezig in het eigen computerarchief. Ook hier wordt nagegaan of en in welke vorm de bestanden van 1971 kunnen worden gebruikt als aanvulling op de gepubliceerde tellingen en/of lichtdrukken. Bij dit onderzoek staat de privacy van de ondervraagden voorop. De gegevens zijn overigens reeds in het verleden geanonimiseerd.

3. Overtypen of automatisch herkennen?

Bij digitalisering moet onderscheid gemaakt worden tussen mediumconversie en inhoudsconversie. Bij mediumconversie wordt een beeld via een *scanner* overgezet op een digitale drager: er ontstaat een digitale kopie of *image*. Bij inhoudsconversie wordt de inhoud van een document in de vorm van cijfers en letters opgeslagen in een structuur die volledige ontsluiting en analyse van de gegevens toestaat.
Uit conserveringsoogpunt biedt mediumconversie van de volkstellingen via *imaging* met indexen op de digitale beelden een oplossing, die vergelijkbaar is met microverfilming, maar die daarenboven enkele voordelen heeft (vinden van informatie, schermpresentatie, afdrukmogelijkheden, distributie). Vanuit het oogpunt van ontsluiting en onderzoek van de volkstellingen is *imaging* echter ontoereikend.
Op dit moment zijn alle gepubliceerde volkstellingen van Nederland via scanning omgezet in digitale images (mediumconversie). Alleen de Volkstelling (inclusief beroepstelling en Woningstatistiek) van 1899 is tevens beschikbaar als database (inhoudsconversie). In de komende jaren zullen echter ook databases van diverse andere tellingen toegankelijk worden gemaakt.
Er bestaan twee manieren om de gegevensinhoud van papier om te zetten naar digitale bestanden: overtypen of met behulp van een programma voor optische tekenherkenning (OCR - *Optical Character Recognition*). Het NIWI is gespecialiseerd in het automatisch herkennen van historische documenten, maar bij tabellen doen zich extra problemen voor ten opzichte van tekst. Bij tabellen is het essentieel dat de gegevens in de juiste rij en kolom terechtkomen.

- De lijnen in de tabellen moeten worden weggefilterd.
- Verwarring van de letters l en O met de cijfers 1 en 0 (en soms ook letter S met cijfer 5).
- Herkenning van de tabelstructuur: scheiding van tabeltitels, opschriften, rij- en kolombeschrijvingen en de eigenlijke inhoud van de tabel.
- Het herkennen van de tabelcellen (rijen en kolommen).
- Herkenning van structuur (hiërarchie) in opschriften van regels en kolommen. De informatie in de tabellen is vaak meerdimensionaal. In de volkstellingen komen in de rijen geregeld vier of vijf geneste niveaus voor (bijv.: 'gemeente > kom > buurt > woningtype'). Ook de kolommen zijn dikwijls genest, maar is deze meer typisch beperkt tot twee of drie niveaus (bijv. 'tijdelijk aanwezigen > mannen').

In het kader van het project is een onderzoek uitgevoerd naar de mogelijkheden van het automatisch herkennen van tabellen. De resultaten hiervan zijn niet alleen toepasbaar op de volkstellingen, maar in het algemeen op informatie die in rijen en kolommen is gerangschikt. Het onderzoek, dat is verricht door Mark Schravesande van de TU Twente, heeft zich vooral gericht op het correct herkennen van de documentstructuur op basis van de uitvoer van OCR-software, die de coördinaten van ieder herkend teken levert. Bij de structurering wordt onderscheid gemaakt tussen rij- en kolombeschrijvingen en de eigenlijke inhoud van de tabel. Er is software ontwikkeld die, met gebruikmaking van een 'tweedimensionale grammatica' de tabelinformatie op een zodanige manier structureert, dat deze automatisch in de juiste rij en kolom wordt geplaatst. Bij het herkennen en structureren van de rij- en kolombeschrijvingen kan de operateur handmatig verbeteringen aanbrengen. Indien de opschriften zich over meer pagina's herhalen, kunnen eenmaal correct herkende en gestructureerde rij- en kolombeschrijvingen worden hergebruikt.
Vervolgens beslist de software op basis van de coördinaten van de tekens welke informatie bij elkaar in welke cel van de tabel behoort. In gevallen waarbij dit ambigu is, worden markeringen geplaatst die een indicatie geven over de aard van de 'twijfel' (bijv.: rij niet zeker, kolom niet zeker, juistheid tekens niet zeker). Bij tests is gebleken is dat de coördinaten van de tekens uit de OCR-uitvoer niet altijd betrouwbaar waren.

4. Mediumconversie door middel van imaging

De 40 microfilms met ruim 42.000 pagina's (twee pagina's per opname) zijn gescand met een Sunrise microfilmscanner. De images zijn in zwart-wit gescand (geen grijswaarden) met een resolutie van 300 dpi (*dots per inch*). De images zijn opgeslagen als TIFF groep 4 (*lossless compression*) en vergen in totaal ca. 3 Gb aan opslag. Op een deel van de images zijn de volgende vormen van digitale beeldverbetering toegepast:

- *cropping:* het verwijderen van zwarte randen rondom de opnamen van de pagina's, die zijn ontstaan bij de microverfilming.
- *noise (speck) removal*: het verwijderen van kleine vlekjes op de afbeeldingen
- *deskewing*: het 'rechtzetten' (zodat de tekstregels horizontaal lopen) van afbeeldingen van pagina's die scheef op de microfilm stonden.

Alle images zijn gecontroleerd op kwaliteit (leesbaarheid). Ook is gecontroleerd op ontbrekende pagina's. Op basis van de kwaliteitscontroles zijn enkele honderden scans extra of opnieuw uitgevoerd. Uitklapkaarten en pagina's met afbeeldingen die in kleur waren gedrukt zijn met een digitale Kontron-camera (in kleur) gescand (ca. 80 kaarten en grafieken).
Voor het opzoeken en raadplegen van de images van de tellingen van 1795-1971 zijn indexbestanden gemaakt. Hierin is onder andere het onderwerp op basis van de titelinformatie van de tabellen, teksthoofdstukken en andere belangrijke boekonderdelen opgenomen. Ook de bibliografische informatie over de boekdelen is in een indexbestand geregistreerd. De bibliografische informatie en de titelbeschrijvingen zijn zoveel mogelijk aan bestaande bestanden ontleend. De deel- en paginanummers van de tabellen zijn handmatig aan de images gekoppeld.
Nagegaan is of het mogelijk zou zijn om de gemeentenamen in de indexen op te nemen. De koppeling van de gemeentenamen met de images bleek echter uiterst arbeidsintensief, omdat per image één of meer (soms tientallen) gemeentenamen kunnen voorkomen. Om deze reden is van deze mogelijkheid afgezien.

5. Inhoudsconversie van de telling van 1899

De tekst uit de inleiding van de Volkstelling 1899 is bij het NIWI via scanning en OCR geconverteerd. Dat is ook gedaan met de kleine tabellen uit de inleiding (van twee pagina's of minder) en met de voorkolom van het rijksdeel van de beroepstelling. Voor het overige zijn alle tabellen van de Volkstelling 1899 (incl. beroeps- en woningtelling) handmatig ingetoetst. Dat is gedeeltelijk gedaan bij het CBS (vestiging Heerlen) en gedeeltelijk, in opdracht van het CBS, bij IVA Data Entry Services BV te Rijswijk (vestiging Heerlen). Bij het handmatig intoetsen heeft in alle gevallen controletoetsing plaats gevonden om het aantal invoerfouten zoveel mogelijk te beperken.
De inhoudsconversie heeft dus grotendeels handmatig plaatsgevonden. Tabel 2 geeft een overzicht van de aantallen verwerkte pagina's van de Volkstelling 1899 (incl. beroeps- en woningtelling). Het totaal aantal ingevoerde pagina's bedraagt ca. 10.000, waarvan de eigenlijke volkstelling er ca. 3000 omvat. Uitgaande van een geschat gemiddeld aantal tekens per tabel-pagina van ruim 1700, is geschat dat het totaal aantal tekens in de tellingen ca. 17 miljoen bedraagt.

Tabel 2: Aantal pagina's Volkstelling 1899 (incl. beroepstelling en woningtelling)

Deel	Volkstelling	Beroepstelling	Woningtelling	Totaal
Inleiding				498
Noord-Brabant	350	521		871
Gelderland	319	625		944
Zuid-Holland	371	889		1260
Noord-Holland	297	703		1000
Zeeland	182	183		365
Utrecht	152	259		411
Friesland	258	575		833
Overijssel	216	515		731
Groningen	184	369		553
Drenthe	112	167		279
Limburg	242	329		571
Het Rijk - Totaal	205	271		476
Het Rijk – Gemeentegrootte-klassen*		959		959
Het Rijk – Woningtelling			174	174
Totaal	**3058**	**6524**	**174**	**9925**

* Niet ingevoerd wegens redundantie

Besloten is om het deel van de beroepstelling, waarin de gegevens zijn gepubliceerd naar gemeente-grootteklassen, niet in te voeren. Met het beschikbaar komen van een bestand met gegevens per gemeente is deze informatie grotendeels redundant. Ook voor controle op de juistheid voegt dat deel weinig toe, omdat er al tal van andere aggregaties zijn, die controle mogelijk maken.

Controles en correcties op de gegevensbestanden

Bij de inhoudsconversie is er naar gestreefd om geen informatie verloren te laten gaan en om de gegevens op een zo 'brongetrouw' mogelijke manier over te nemen uit de publicaties. In principe is alle informatie uit de tabellen en toelichtende teksten gedigitaliseerd. De gebruiker heeft bovendien de digitale images als controlemiddel.
Na voltooiing van de data-invoer van de tabellen van de Volkstelling 1899 zijn controles op de juistheid van de gegevens in de database uitgevoerd. Belangrijkste instrument hierbij vormde het vergelijken van in de bron gegeven totalen met berekende totalen. Er kunnen zich verschillende typen van fouten voordoen:

- Invoerfouten: deze zijn gecorrigeerd. Omdat dubbel is getoetst, zijn slechts weinig data-entryfouten gevonden.
- Fouten doordat de gedrukte informatie onleesbaar is: waar mogelijk kunnen de waarden worden afgeleid uit andere gegevens in de bron. In de volkstelling van 1899 komt dit in beperkte mate voor. Deze fouten zijn gecorrigeerd.

- Bronfouten: berekende totalen kunnen afwijken van gegeven totalen door druk- of optelfouten. Controle heeft plaats gevonden door totaliseren van rijen en kolommen en vergelijking met rij-, respectievelijk kolomtotalen. Deze fouten zijn niet gecorrigeerd. Wel is er een bestand beschikbaar met een overzicht van geconstateerde bronfouten en, waar mogelijk, suggesties voor correctie.
- Ook in voetnoten opgenomen aantallen personen bleken effect te hebben op ogenschijnlijke fouten. Noten bij de tabellen zijn aangebracht en gedeeltelijk herleid tot extra tabelkolommen (in sommige gevallen zijn in de oorspronkelijke publicaties weinig voorkomende getallen in voetnoten vermeld om ruimte te sparen).

De meest uitvoerige foutenanalyse is gemaakt van het zogenaamde 'Rijksdeel' van de beroepstelling van 1899, dat uit 6424 regels (= records) bestaat. Deze analyse is grotendeels gemaakt door Tom Vreugdenhil van het CBS, aan wie de nu volgende bevindingen zijn ontleend. In elke regel komen zes totaalcijfers voor. Deze totalen werden vergeleken met de optelling van de bijbehorende cellen, na correctie met de gegevens in de voetnoten. Er waren 80 regels waarin één of meer verschillen werden gemeld. Voor deze regels werden de resultaten van de data-entry vergeleken met de digitale afbeeldingen beelden van de gedrukte pagina's op CD-ROM.
Uit deze vergelijking bleken slechts twee data-entry-fouten; in beide gevallen betreft het een verschuiving van het type " 0 2 2 0 0 " in plaats van " 0 0 2 2 0 ". Daarnaast werden in de laatste twee regels in twee cellen het begincijfer vermist; het betreft de cellen die in de gedrukte publicatie links op de rechter pagina staan, direct tegen de band. Aangenomen mag worden, dat de data-typistende cijfers niet goed konden zien, omdat ze in de band verdwenen waren.
Van de overige 76 regels met verschilmeldingen zijn met name de grotere verschillen veroorzaakt door duidelijke tel- of zetfouten en dus gemakkelijk te herstellen. Andere gevallen zijn veel lastiger te detecteren; op zijn minst is dan ook een analyse nodig van de optellingen in de kolommen.
Controle van de optellingen in kolommen is wat ingewikkelder, maar niet onmogelijk. Hierbij kon gebruik worden gemaakt van de wijze waarop de beroepen waren ingedeeld. Er bleken drie plekken met verschilmeldingen over vrijwel de gehele regel. In deze gevallen bleken in de gedrukte publicatie verkeerde letters te zijn gebruikt voor de positie in het beroep, die was aangeduid met één van de eerste vier letters van het alfabet. Een zeer evident geval was de volgorde C, D, D, D in het totaal van een categorie waar normaal A, B, C, D staat.
Ook hier geldt, dat een aantal verschilmeldingen duidelijk wordt veroorzaakt door tel- of zetfouten. Over het precieze aantal verschillen is moeilijk iets te zeggen, omdat een verschil ook weer kan doorwerken in meer geaggregeerde totalen.
Van de moeilijker gevallen is er één wat dieper uitgezocht. Het betreft een geval waarbij in de tabel van enkele provincies beroepen zijn gebruikt die in het totaal van het Rijk niet voorkomen en daar kennelijk bij andere beroepen zijn geteld. Vergelijking van de tabellen van de provincies, de gemeentegroottegroepen en het Rijk maakte in dit geval duidelijk waar de fout zat.
Het maken van een consistent geheel, waarbij alle totalen kloppen, zowel binnen de onderdelen, maar ook voor het totaal van de provincies met elk van de cijfers van het Rijk bleek een opgave die buitengewoon veel tijd zou kosten (en ook nooit tot de projectopzet heeft behoord). Veel verschillen kunnen worden opgespoord, maar er is geen garantie dat alle problemen kunnen worden opgelost, niet altijd kan worden achterhaald of er sprake is van een foute optelling of van een zet- of drukfout in een niet geaggregeerd getal.
Ten aanzien van de Inleiding 1899 is geprobeerd het uiterlijk van de elektronische tekst zo goed mogelijk te laten lijken op de oorspronkelijk gedrukte pagina's. Bij de lijsten en tabellen is dit principe minder strikt toegepast. Hierbij stonden twee uitgangspunten centraal: ten eerste het algemene principe dat geen informatie verloren mocht gaan en ten tweede het gebruikersgemak bij verdere verwerking (bijvoorbeeld analyse in een spreadsheet-programma). Zo zijn bijvoorbeeld tabellen en lijsten, die in het boek in kolommen waren gezet, doorgaans omgezet naar één kolom. Oplossingen werden gevonden voor herhalingen (die met aanhalingstekens waren aangegeven) en voor cellen die waren samengevoegd (accolades in de gedrukte tabellen). Voetnoten in tabellen zijn in afzonderlijke kolommen opgenomen. Verduidelijkingen en toevoegingen (bijvoorbeeld van opschriften en titels) zijn tussen teksthaken geplaatst.
In de inleiding op de Volkstelling 1899 is de variëteit aan tabellen zeer groot en blijkt ook de diversiteit aan fouten groot te zijn (NB: het gaat hier om fouten die door de samenstellers van de telling van 1899 gemaakt zijn, niet om data-entry fouten!). Dit hangt ongetwijfeld samen met het feit dat in deze relatief kleine, samenvattende tabellen, zeer uiteenlopende berekeningen en bewerkingen zijn uitgevoerd.

De beroepenclassificatie van 1899

Voor de classificatie van de beroepen in 1899 is aanvankelijk uitgegaan van bijlage II uit de Inleiding. Verondersteld werd dat deze classificatie alle beroepen uit de telling zou omvatten. In de classificatie worden vier hiërarchische niveaus onderscheiden. De beroepen uit deze lijst zouden worden gekoppeld aan de cijfers van de beroepstelling. Bij controles bleken echter aanzienlijke verschillen in beroepsomschrijvingen te bestaan tussen de classificatie en de formuleringen in de twaalf delen van de beroepstelling. Hierop zijn ook de beroepsomschrijvingen uit de beroepstelling van het Rijk als geheel gedigitaliseerd via scanning en OCR, alsmede de omschrijvingen uit bijlage III van de Inleiding. Na analyse van de verschillende lijsten is die uit het Rijksdeel van de beroepstelling genomen als groslijst voor de koppeling van de beroepenclassificatie met de cijfers in de tabellen. Toch werden tijdens de invoer nog steeds afwijkingen gevonden. Iedere afwijking van een beroepstitel die niet duidelijk een drukfout betrof, werd in het bestand geregistreerd. Op het totaal van ca. 100.000 records van de beroepstelling 1899 bleken ruim 2.500 beroepstitels niet in de groslijst voor te komen. Deze varianten zijn achteraf afzonderlijk behandeld en alsnog geklasseerd. Het resultaat is verwerkt in de beroepenclassificatie voor 1899. In een afzonderlijk bestand zijn de beslissingen met betrekking tot de beroepenclassificatie vastgelegd.

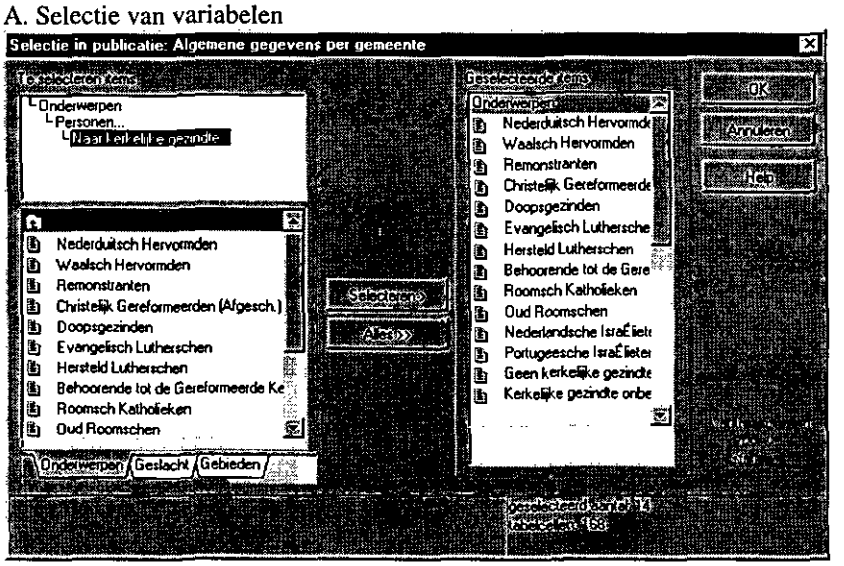
6. Ontsluiting

De databestanden van de Volkstelling 1899 zijn ontsloten met StatLine, via welk systeem ook andere CBS-statistieken toegankelijk zijn. StatLine maakt het mogelijk dat de gebruiker zelf zijn tabellen samenstelt uit de beschikbare rij- en kolomvariabelen. De paginanummers van de oorspronkelijke tabel in de boeken zijn in de tabellen weer te geven, zodat de gebruiker de gegevens kan controleren op de meegeleverde images.

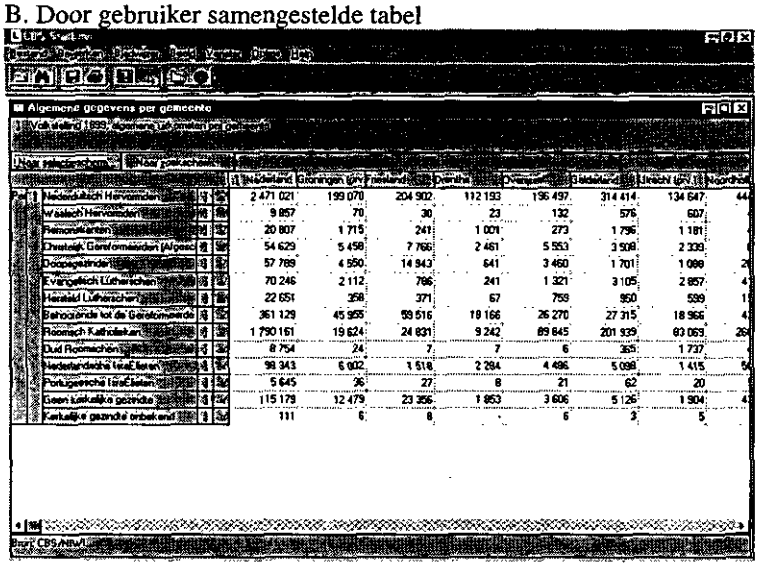
A. Selectie van variabelen	B. Door gebruiker samengestelde tabel
C. Weergave als grafiek	D. Weergave als kaart

Figuur 1: Manipulatie van de gegevens van de Volkstelling 1899 in CBS StatLine

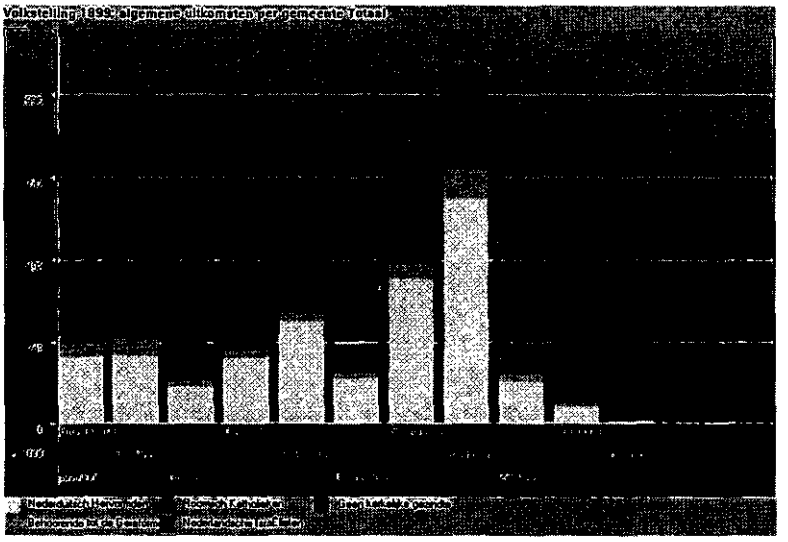
A. Selectie van variabelen



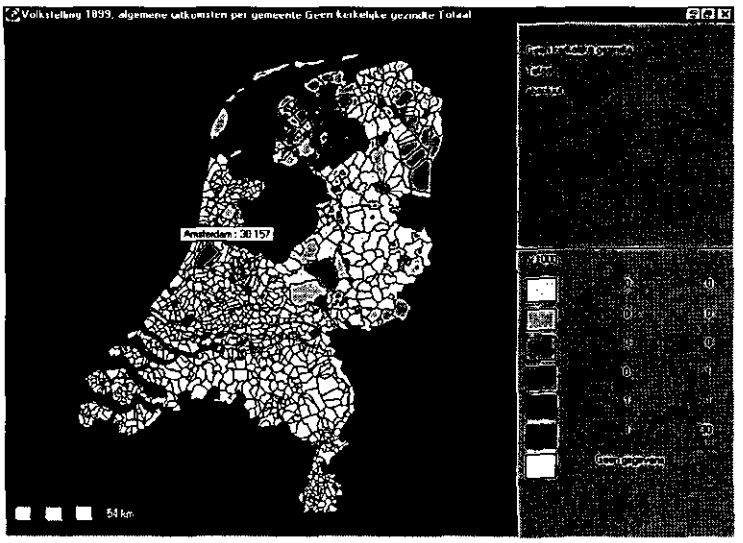
B. Door gebruiker samengestelde tabel



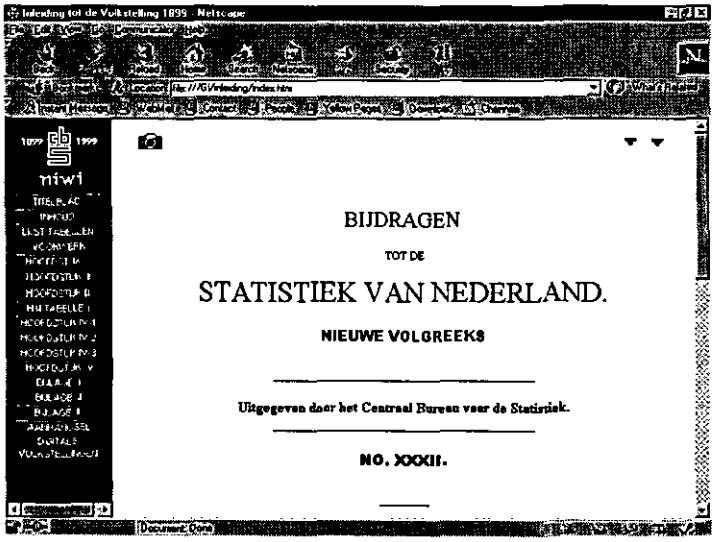
C. Weergave als grafiek



D. Weergave als kaart

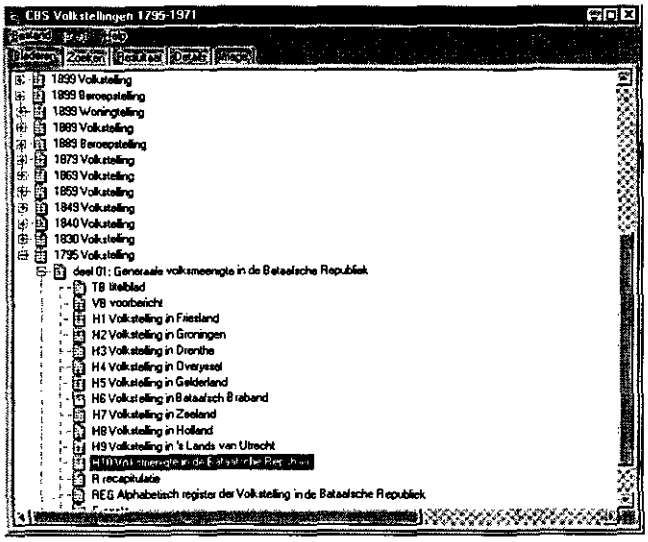


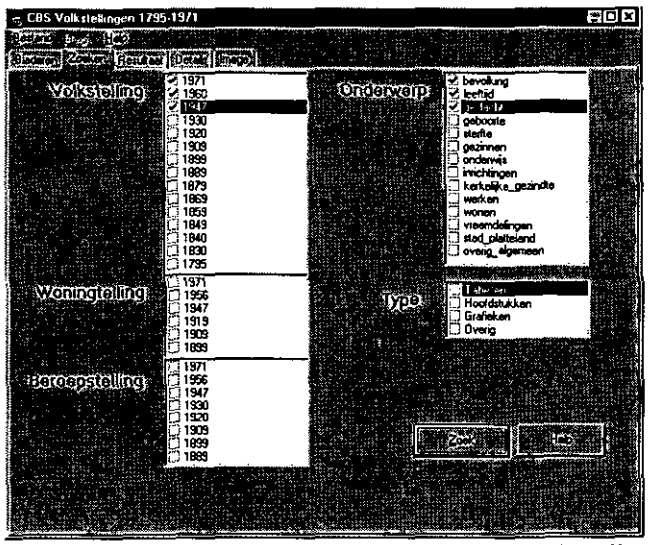
StatLine is zowel beschikbaar op CD-ROM als op het World Wide Web. Door de gebruiker gemaakte data-selecties kunnen eenvoudig kunnen worden opgeslagen om verder te worden bewerkt in een spreadsheet of statistisch pakket. De gegevens kunnen ook als grafiek of kaart (op gemeenteniveau) worden weergegeven in de CD-ROM versie van StatLine.
De Inleiding 1899 is ontsloten via een speciale Website. Deze is toegankelijk op het World Wide Web (op adres www.volkstelling.nl), maar ook op CD-ROM. Bij het maken van de Website is ernaar gestreefd om de presentatie van de tekst het uiterlijk te geven van de oorspronkelijke publicatie. De gebruiker kan dit controleren, omdat van iedere pagina ook een digitale afbeelding aanwezig is. De tekst en tabellen zijn geheel doorzoekbaar. De tabellen uit het inleidende deel volgen de oorspronkelijke vormgeving niet nauwgezet. Uitgangspunt voor de presentatie van de tabellen was het gemak voor eventuele verdere verwerking. De tabellen kunnen als afzonderlijke bestanden worden opgeslagen op een eigen schijf en vervolgens verder verwerkt in een spreadsheet-programma.



Figuur 2: De homepage van de Inleiding op de telling van 1899 (www.volkstelling.nl)

Voor het opzoeken en raadplegen van de images van de tellingen van 1795-1971 is een apart zoeksysteem gemaakt. Er kan gezocht worden op het jaar van de telling, het soort of deel van de telling en op onderwerp. Het is ook mogelijk om door de images te bladeren.





Figuur 3: Bladeren en zoeken door de images van de volkstellingen

7. Conclusies

Het project Digitalisering Nederlandse Volkstellingen 1795-1971 heeft in ruim twee jaar tijd de eerste concrete producten opgeleverd in de vorm van een website en twee sets CD-ROM's. De belangstelling van de doelgroep voor de resultaten van het project is groot. Het sociaal- en economisch historisch onderzoek wordt door het project gestimuleerd. In het najaar van 1999 heeft een symposium plaatsgevonden, waarbij de Volkstelling van 1899 het uitgangspunt vormde. Het project heeft ook veel kennis opgeleverd over de opzet en aanpak van een gecompliceerd project, waarin zowel medium- als inhoudsconversie een rol spelen. Enkele belangrijke lessen uit het project zijn:

- Microverfilming is een zeer zinvolle en efficiënte tussenstap bij omvangrijke digitaliseringsprojecten. Niet alleen wordt een extra *hardcopy backup* gemaakt, ook het werkproces verloopt efficiënter dan bij scanning van papier. Bij verwerking van zwart-wit drukwerk treedt geen noemenswaardig kwaliteitsverlies op. De microverfilming dient wel uiterst zorgvuldig en goed afgestemd op de scanning te geschieden. De verkleiningsfactor moet zo klein mogelijk zijn en de film moet contrastrijk zijn. Voor automatische scanning moeten de rolfilms zo lang mogelijk zijn en mogen deze geen lassen bevatten. De scanner ondervindt door lassen problemen met het automatisch vinden van de beelden op de film.

Controle op kwaliteit en volledigheid van de opnamen is een vereiste. In het volkstellingenproject hebben correcties en aanvullende scans (enkele honderden opnamen) een veelvoud van de benodigde tijd van het automatisch scannen van het volledige filmmateriaal gekost (ruim 20.000 beelden).

- Voor inhoudsconversie van omvangrijke tabellen in grote hoeveelheden is nog geen produktierijpe software voor handen. Onderzoek en ontwikkeling hebben wel verbetering gebracht in de beschikbare hulpmiddelen voor het automatisch structureren van cijfermateriaal in rijen en kolommen, maar deze leveren nog geen overtuigende efficiency-winst. In principe is de aanpak, waarbij gebruik wordt gemaakt van de XY-coördinaten van OCR-uitvoer, wel veelbelovend.
- Controle en correctie van de gedigitaliseerde informatie is de meest arbeidsintensieve fase van een conversieproject. Ook in dit project was dit het geval. In dit project waren de volgende onderdelen veel tijdrovender dan voorzien:
  - Controle op bronfouten; hoewel inconsistenties tussen met de computer berekende totalen en in de oorspronkelijke bron afgedrukte (vroeger handmatig berekende) totalen indicaties opleveren voor optredende fouten, is de oorzaak daarvan niet altijd te achterhalen en is correctie daardoor niet verantwoord.
  - Het opstellen en toepassen van een consistent, hiërarchisch classificatiesysteem op historische beroepen. De beroepstitels en -indelingen van 1899 blijken niet volledig en niet geheel consistent te zijn.
  - Het controleren en indiceren van de images aan de hand van bestaande overzichten van tabellen en inhoudsopgaven. Door inconsistenties in de inhoudsopgaven en lijsten van tabellen en door veranderingen in terminologie in de loop der tijd is het geautomatiseerd creëren van toegangen op de images problematisch.

De resultaten van het project rechtvaardigen een vervolg. Bij dit vervolg is het wenselijk aandacht te besteden aan de volgende zaken:

- Ontsluiting en toegankelijkstelling van een groter aantal historische volkstellingen; een deel van de data-entry heeft al plaats gevonden, maar de ruwe data zijn zonder aanvullende ontsluiting niet bruikbaar voor onderzoekers.
- Digitale publicatie van de Volkstelling 1960; het digitale basismateriaal van deze telling, zij het met lacunes, is gedeponeerd bij het Steinmetzarchief van het NIWI; voor deze telling kan als publicatieformaat worden aangesloten bij de gedetailleerde, handgeschreven tabellen in het archief en de bibliotheek van het CBS. Aan de hand van dit materiaal kunnen de ontbrekende gegevens in de digitale telling worden aangevuld.
- Digitale publicatie van de Volkstelling 1971, waarvan het (anonieme) basismateriaal zich in het digitale archief van het CBS bevindt. Ook voor deze telling geldt dat een publicatievorm moet worden gevonden, die recht doet aan de privacy van de getelde bevolking.

Bij een vervolgproject moet vooral duidelijk worden bepaald, in hoeverre verschillende tellingen onderling vergelijkbaar zijn. Het onlangs afgeronde project laat zien, dat juist redigeerwerk (correctie, classificatie) zeer moeilijk en arbeidsintensief is. Gemeentelijke herindelingen, veranderende definities en uiteenlopende classificaties zullen een volledig consistente set digitale volkstellingen tot een illusie maken. Het zal een schone taak voor het historisch onderzoek zijn om na te gaan in hoeverre de volkstellingen in de loop der tijd vergelijkbaar gemaakt kunnen worden. En dan te bedenken dat de volkstellingen alleen maar het begin vormen van de onafzienbare hoeveelheden historische statistieken die in Nederland bewaard zijn gebleven...

Bijlage 1. Medewerkenden en dankwoord

Dit project is tot stand gekomen dankzij de medewerking van velen, in velerlei vorm. Voor die hulp en voor het enthousiasme waarmee, zowel binnen als buiten CBS en NIWI, aan het project werd meegewerkt, willen wij hen hartelijk bedanken. We noemen in het bijzonder:

CBS:
dr. Jacques van Maarseveen (projectleider)
dr. Tom Vreugdenhil (materiaalselectie, controle data Volkstelling 1899 en ontsluiting data Volkstelling 1899 via StatLine)
drs. Jan Jonker (inhoudelijke begeleiding data-entry, codering beroepen, structurering tabellen Inleiding Volkstelling 1899)
mevr. Jolanda Rikers (marktonderzoek)
Sjef Römkens, Ad Bemelen, Jan Daamen, mevr. Riny Lauvenberg, mevr. Guus Crol (organisatie en begeleiding data-entry)
Jo Florie (data-entry programmatuur)
mevr. Wilma Schusser, mevr. Elly Moonen en vele collega's van de sectoren HIH, HVV, SIV en GAB (data-entry)

NIWI:
dr. Peter Doorn (projectleider)
drs. Hans van Mourik (indexen, onderzoek Volkstelling 1960, ontsluiting images Volkstelling 1795-1971)
drs. Mark Schravesande (Technische Universiteit Twente; onderzoek tabelherkenning)
drs. Dagmar Stiebral (indexen, controle images, Website Inleiding Volkstelling 1899)
drs. Rubrecht Zaat (scanning, onderzoek conversie lichtdrukken)
Annuska Graver (marktonderzoek)
Roselyne van der Heul (codering beroepen, correctie en structurering Inleiding Volkstelling 1899)
drs. Esmeralde Marsman (codering beroepen)
drs. Marjana Rhebergen (correctie en structurering Inleiding Volkstelling 1899)
drs. René van Horik (vooronderzoek, optische tekenherkenning)
drs. Jurriën de Jong (Universiteit Leiden; codering beroepen)
drs. Berry Feith (controle en correctie images, HTML-markup)
drs. Niek van Baalen (zoeken web-site Inleiding Volkstelling 1899)
Bram Buitendijk (zoeken web-site Inleiding Volkstelling 1899)

Bibliotheken:
De volgende bibliotheken stelden exemplaren van hun volkstellingsboeken ter beschikking voor verfilming en digitalisering:
Centraal Bureau voor de Statistiek
Universiteit Leiden
Rijksuniversiteit Groningen
Internationaal Instituut voor Sociale Geschiedenis

Data Entry:
IVA Data Entry Services BV, Rijswijk/Heerlen
Franco Bonafini (data-entry programmatuur)
mevr. Toos Beckers (supervisie)

Microverfilming:
Microformat, Lisse

Financiering:
Innovatie Wetenschappelijke Informatievoorziening, SURF
Gebiedsbestuur Maatschappij- en Gedragswetenschappen / Wetenschappelijk Statistisch Agentschap, NWO
Nederlands Instituut voor Wetenschappelijke Informatiediensten, KNAW
Centraal Bureau voor de Statistiek

Bijlage 2. Bestelinformatie

De Nederlandse Volkstellingen 1795-1971 zijn digitaal beschikbaar op cd-rom en voor een deel op het Internet.

De cd-rom uitgave omvat twee sets cd-rom's:

- Set 1: Data en publicatie volkstelling 1899. Twee cd-rom's. ISBN 90.6861.176.3, fl. 129,00. Deze uitgave bevat de volledige tekst van de Inleiding op de Volkstelling van 1899, alle gepubliceerde tabellen als StatLine-databases en digitale afbeeldingen van alle ca. 10.000 pagina's van de oorspronkelijke gedrukte uitgave. Met het bijgeleverde Statline-programma kan de gebruiker tabellen samenstellen voor een zelf gekozen selectie van gegevens.
- Set 2: Publicaties volkstellingen 1795-1971. Vijf cd-rom's, ISBN 90.6861.177.1, fl. 249,00. Deze uitgave bevat digitale afbeeldingen van de ca. 42.500 pagina's van alle gedrukte publicaties van de Volkstellingen tussen 1795 en 1971.

De cd-rom's zijn te koop bij de boekhandel of rechtstreeks te bestellen bij:
Uitgeverij Stichting beheer IISG, Cruquiusweg 31, 1019 AT Amsterdam, Tel. (020) 668 58 66, fax (020) 665 64 11

De homepage van de Internet-publicatie is www.volkstelling.nl. Hier bevindt zich de volledige tekst van de Inleiding tot de Volkstelling van 1899 met bijbehorende tabellen en grafieken. Alleen de Web-versie van deze Inleiding is volledig doorzoekbaar.
De StatLine-publicatie van de tabellen van de Volkstelling 1899 is tevens bereikbaar via de home page van het CBS www.cbs.nl.

VAN NU EN STRAKS - EN VEEL LATER

De stand van zaken met betrekking tot behoud en toegankelijkheid van informatiedragers

Sjaak Hubregtse

In de conferentiebrochure wordt gesteld dat duurzaamheid op drie manieren een rol speelt: duurzaamheid van digitale informatie, van huidige ontwikkelingen en van bestaande instituties. Maar echt vergelijkbaar zijn die drie gebieden niet: het behoud van digitale informatie - of beter: van informatie in het algemeen - is een actueel probleem waarvoor nú concrete oplossingen worden gezocht en soms ook gevonden. Alleen op dit eerste aspect zal ik wat uitvoeriger ingaan; ik doe dat niet als IC-technoloog, maar vanuit het perspectief van de informatiemanager. De andere twee aspecten nodigen veeleer uit tot koffiedikkijken en cultuurfilosofische beschouwingen. Ik zal me daar nauwelijks aan wagen.

1 Probleem en bewustwording[1]

Informatie kan niet in een duurzame vorm bestaan als ze niet op of in enig medium is opgeslagen - ongeacht of dat medium nu klei is, een plankje, papyrus, perkament, papier, vinyl, celluloid of een of andere synthetische stof zoals gebruikt voor een hard disk, CD- ROM, et cetera. Als de fysieke basis verloren gaat, gaat daarmee ook de opgeslagen informatie verloren. Dat is, nu meer dan ooit, het probleem.
Dit probleem is overigens geen vanzelfsprekenheid, en ook niet van alle tijden. Bekijken we de complete informatieketen, waarin 'behoud' of 'bewaring' een schakel is, dan kunnen we makkelijk vaststellen dat niet elke schakel uit de aard der zaak een probleem vormt. Als we vervolgens de eeuwenlange historische ontwikkeling van de informatie-overdracht bezien, dan constateren we dat het behoud van een collectie beslist niet altijd zo'n dominant probleem heeft gevormd. Daar zijn twee eenvoudige verklaringen voor. Eén: tot betrekkelijk kort geleden - zeg begin twintigste eeuw - bestond er uitsluitend informatie in geschreven of gedrukte vorm, en tot iets langer geleden - zeg begin negentiende eeuw - had de drager van die informatie, papier of perkament, een zeer grote duurzaamheid, uit te drukken in eeuwen. Twee: door het kleine aantal potentiële en reële gebruikers èn de beperkte toegankelijkheid lag de gebruiksfrequentie zeer laag en had het materiaal dus relatief weinig te lijden.

Het probleem is nu actueel en de belangstelling ervoor schijnt van recente datum te zijn. Dat laatste is waar en niet waar; ik verklaar mij nader d.m.v. een zeer kort historisch overzicht.
John Feather, van de Engelse Loughborough University, schreef in 1991: 'Preservation has become one of the less predictable fashions in librarianship in the last decade of the twentieth century'. (Feather 1991, zie ook Feather e.a. 1996)
Andere auteurs rond 1990 noemden preservation 'the glamour issue of the library world' of schreven dat 'preservation may understandably be seen as somewhat of a newcomer'. In werkelijkheid schreef de bibliothecaris van de Library of Congress al in 1876, het zogenaamde *annus mirabilis* in de Amerikaanse bibliotheekgeschiedenis, 'there is no subject more important in the administration of a public library than the binding and preservation of the volumes'.
In 1887 werd in St. Gall, Zwitserland, een internationaal congres georganiseerd over behoud en restauratie van manuscripten; in Washington verscheen in 1909 een studie met de titel *Durability and economy in papers for permanent records.*
Die belangstelling, nu een eeuw geleden, voor het onderwerp, en dan met name de duurzaamheid van het papier, is niet verbazingwekkend. De eeuw tussen ca. 1850 en 1950 staat bekend als de periode waarin het slechtste papier uit de hele boekgeschiedenis werd gefabriceerd (namelijk: van huis uit verzuurd), en deskundigen kenden natuurlijk de gevaren daarvan. Veel merkwaardiger is het dat die belangstelling in het begin van de twintigste eeuw weer geheel verdwijnt, en nòg vreemder is, dat het tot het eind van die eeuw duurt voor ze weer opbloeit, terwijl er toch enkele decennia eerder twee schokkende gebeurtenissen plaatsvonden.

Namelijk: In de jaren '60 wordt men zich vooral in grote Amerikaanse bibliotheken opnieuw bewust van het verzuringsprobleem: het verval heeft intussen zodanig om zich heen gegrepen dat grote delen van collecties niet meer raadpleegbaar zijn. In 1963 begroot de New York Public Library dat het 12 miljoen dollar zal kosten om alleen nog maar het hoogst nodige te doen. In de jaren '70 schat men dat van de wetenschappelijke bibliotheekcollecties in New York City 30-50% in ernstige staat van verval verkeert. Om hele collecties te behouden, ontwikkelt men nieuwe technieken, zoals grootschalige ontzuring en paper splitting.

Het tweede en nog beslissender signaal wordt gevormd door de overstroming van Florence in november 1966. De rivier de Arno treedt buiten zijn oevers; o.a. de door Michelangelo ontworpen Biblioteca Medicea Laurenziana wordt getroffen en bijna een half miljoen manuscripten en zeldzame boeken lopen ernstige waterschade op.
Hoewel de ramp kwantitatief in feite minder ernstig is dan het wereldomvattende verzuringsprobleem, reageert het publiek wereldwijd geschokt en realiseert zich de kwetsbaarheid van het culturele en intellectuele erfgoed; de Unesco start een internationale reddingsoperatie. Ook hier worden nieuwe technieken ontwikkeld, bijvoorbeeld vriesdrogen.

Toch duurt het ook nu nog enkele decennia voor de zorg om het behoud van kennis en cultuur wordt geïnstitutionaliseerd. We zien dan het volgende gebeuren:

1984 In Washington wordt opgericht de Commission on Preservation and Access (CPA).
1984 Officieel begin van het IFLA-kernprogramma Preservation and Conservation (PAC); sinds 1992 is de Franse Bibliothèque Nationale het centrum hiervan.
1992 Unesco start het *Memory of the World* Programma, 'with the twofold purpose to safeguard and promote the endangered world documentary heritage'.
1994 De European Commission on Preservation and Access (ECPA) wordt opgericht; ze zetelt in het gebouw van de KNAW te Amsterdam.

Deze en vergelijkbare instituties besteden sindsdien op vele manieren aandacht aan het probleem, onder andere door congressen, symposia en conferenties te organiseren - tot op de dag van vandaag.

2 Toegankelijkheid

Hierboven heb ik geschetst wat het probleem is. Op het gevaar af naar de bekende weg te vragen, stel ik nu toch de vraag: Waarom is dat eigenlijk een probleem?
Natuurlijk is het geen probleem omdat behoud een doel op zich zou zijn, want dat is het niet.

Behoud is 'slechts' een middel om het werkelijke doel te realiseren: toegankelijkheid. In tegenstelling tot de hierboven genoemde situatie van enkele eeuwen gelden, toen vele bibliotheken een half-gesloten of zelfs museaal karakter hadden en collectiebehoud mogelijk als primair doel gezien kon worden, heeft nu toegankelijkheid het primaat.
De behoefte aan toegankelijkheid is de enige reden en rechtvaardiging voor behoud. Wat ontoegankelijk wordt bewaard, wordt zinloos bewaard. Daarom is het ook zinloos om over *behoud* te spreken zonder het tegelijk over *toegankelijkheid* te hebben; daarom kom je die twee woorden ook altijd samen tegen, of het verwante begrippenpaar *bewaring en beschikbaarstelling,* of in het Engels: *preservation and access.*

Toegankelijkheid speelt overigens, net als duurzaamheid, op drie manieren een rol. Naast het tot nu toe genoemde 'technische' aspect - toegang tot de informatie op de drager - is er het 'ideologisch-juridische' en het 'menselijke'.
Het ideologisch-juridische aspect confronteert ons met spanningsvelden tussen publiek en privaat domein, tussen informatierecht en intellectueel eigendomsrecht, et cetera. In Nederland houdt o.a. het Instituut Rathenau zich bezig met de publiek/privaat-kwestie[2]; aan het andere vraagstuk werd in januari jongstleden een driedaags congres gewijd in Krakau, Polen[3], en ik zal me hier met geen van beide bemoeien.
Met 'het menselijk aspect' doel ik op de vraag hoe toegankelijk insitituties en gebouwen zijn, hoe hoog of laag de fysieke en psycholgische drempels zijn die voor de informatiezoeker worden opgeworpen. In Paragraaf 7 komt dit kort ter sprake.

3 Concrete problemen en oplossingen

Diverse instellingen - o.a. de European Commission on Preservation and Access en andere hierboven genoemde - houden zich nu onder meer bezig met dezelfde problematiek als die van honderd jaar geleden, namelijk: verzuurd papier. Maar in de tussentijd zijn er vele nieuwe soorten informatiedragers bijgekomen, die alle hun eigen kwaliteiten en problemen met zich mee brachten.
Magneetbanden bijvoorbeeld (in de VS zijn er onafzienbare archieven mee gevuld, waarbij het bijvoorbeeld om oral history gaat, maar ook om via ruimtevaart verkregen gegevens over de aarde en andere hemellichamen) zijn al na vijf jaar geen betrouwbaar opslagmedium meer: er kan een proces beginnen waarbij de magnetische laag loslaat van de drager. Ook de nog gave exemplaren zijn moeilijk raadpleegbaar, omdat afspeelapparatuur niet of nauwelijks verkrijgbaar is.[4]
De levensduur van de gewone audio-CD en CD-ROM staat nog niet vast, maar is vrijwel zeker zeer beperkt. Amerikaanse laboratoria vrezen dat een CD-ROM van gemiddelde kwaliteit niet langer dan tien jaar meegaat - onder gunstige bewaaromstandigheden.
Veel oude, nitraathoudende films zijn intussen veranderd in een klomp chemisch afval, en onherroepelijk verloren.
Ook de levensduur van onze huis-tuin-en-keuken videoband moeten we uitdrukken in één, hoogstens twee decennia. Na een beperkt aantal jaren niet-afdraaien begint een verklevingsproces en is de band waardeloos.
Digitale opslag moeten we vooralsnog tot de minst betrouwbare media aller tijden rekenen. Bestanden die zijn vastgelegd op CD-ROM van absolute topkwaliteit gaan mogelijk zo'n vijftig jaar mee - maar er is geen enkele garantie dat de benodigde hard- en software dan nog beschikbaar zijn[5].

Hieronder volgen nu de belangrijke media papier, microfilm/fiche en digitale opslag, met hun voor- en nadelen, problemen en alternatieven: recente projecten en praktijkervaringen.

Papier > microverfilming

Ik begin dicht bij huis, met het Nederlandse project *Metamorfoze*, dat op initiatief van het Ministerie van Onderwijs, Cultuur en Wetenschappen van start ging in 1997, en wordt gecoördineerd door het Bureau Conservering Bibliotheekmateriaal van de Koninklijke Bibliotheek. Het programma richt zich op de conservering van handschriften, boeken, kranten en tijdschriften van Nederlandse origine, die aanwezig zijn in bibliotheken met een bewaarfunctie. Doel is, deze materialen voor verder verval te behoeden, en tevens de informatie toegankelijk te houden. Wie zich deze dubbele taak stelt, moet kiezen uit drie alternatieven: hij kan het papier ontzuren, de inhoud digitaliseren, of de inhoud vastleggen op microfiche of -film. *Metamorfoze* heeft gekozen voor de laatste mogelijkheid, en in een toelichtende brochure wordt dat als volgt beargumenteerd.

Waarom microverfilming?

Voor de eerste vier jaar van Metamorfoze is gekozen voor overzetten van het origineel
op microfilm en nog niet voor ontzuring of digitalisering. Aan ontzuring en digitalisering als conserveringsmethoden kleven op dit moment nog bezwaren. Ontzuring kan er weliswaar voor zorgen dat de levensduur van een document verlengd wordt, maar de kwaliteit van het papier wordt er niet door verbeterd. Bovendien wordt er op internationaal niveau nog onderzoek gedaan naar de neveneffecten van ontzuring.
Digitalisering biedt vooral voordelen bij het toegankelijk maken en snel beschikbaar stellen van de opgeslagen informatie. Nadeel is echter dat ten aanzien van de houdbaarheid van digitale media nog veel onzekerheden bestaan.
Een microfilm die in het kader van Metamorfoze wordt gemaakt, gaat minstens 200 jaar mee en voldoet aan de kwaliteitseisen die noodzakelijk zijn voor optimale digitalisering in de toekomst.

Het is de bedoeling dat een en hetzelfde boek nooit meer dan één keer wordt geconverteerd, in binnen- noch buitenland. Daarom wordt elke titel aangemeld bij EROMM, European Register of Microform Masters. Dit register telt momenteel enkele miljoenen titels van boeken en tijdschriften; er kunnen kopieën van de microvorm worden besteld, maar ook printing on demand is mogelijk. EROMM is gevestigd in de Universiteit van Göttingen; voor meer informatie: http://www.gbv.de/help/du/eromm_obn.shtml(zie ook Schwartz 1996)

Papier: massa-ontzuring

Sinds jaren wordt er internationaal geëxperimenteerd met diverse, technisch sterk verschillende, massa-

ontzuringsmethoden. De bekendste daarvan zijn DEZ, Wei T'o, Sablé, FMC, Battelle en Bookkeeper. Voor de respectievelijke technische specificaties is dit niet de juiste plaats (zie daarvoor Porck 1996, die overigens de FMC-methode niet bespreekt). Feit is, dat de DEZ methode inmiddels is afgevallen, wegens de grote risico's (explosiegevaar) voor boek en mens. Van de resterende methoden lijken Battelle en vooral Bookkeeper thans letterlijk en figuurlijk de beste papieren te hebben, mede op grond van TNO-rapporten (Havermans 1996, 1997). Het proces is effectief en chemisch volstrekt risicoloos: roken en open vuur kunnen in de werkruimte geen kwaad.[6] Nadeel is dat sommige boeken die door verzuring sterk verzwakt zijn, mechanisch enigszins te lijden hebben: ze worden tijdens hun verblijf van ca 25 minuten in het ontzuringsbad continu bewogen.

De Library of Congress gebruikt de methode op relatief grote schaal: thans zijn enkele honderdduizenden boeken behandeld, en in 1998 werd een nieuw vierjarig contract gesloten. Nederlandse referenties zijn o.a. de Rijksarchiefdienst, het Algemeen Rijksarchief en de Koninklijke Bibliotheek, alle te Den Haag.
Bij dit alles komt nog, dat zelfs een in alle opzichten volmaakt ontzuringsproces dikwijls slechts in theorie een oplossing zal zijn. Het zet alleen het interne afbraakproces stil, maar het boek wordt er niet sterker door dan het was, en blijft dus veelal ongeschikt voor raadpleging. In de praktijk is ontzuring alleen afdoende als we te maken hebben met een boek dat van huis uit weliswaar verzuurd is, maar nog in goede conditie: preventieve ontzuring. Een en ander leidt ertoe dat uit de praktijk soms het sombere vermoeden klinkt dat het met alle ontzuringsexperimenten binnen afzienbare tijd zal zijn afgelopen, tenzij er heel snel belangrijke ontwikkelingen plaatsvinden. Opmerkelijk in dit verband is, dat zelfs de firma Preservation Technology (nog) niet overtuigd is van de commerciële overlevingskansen van haar veelgeprezen product Bookkeeper[7].

Papier > digitaal

Hoewel er, zoals de Metamorfozefolder zegt, 'ten aanzien van de houdbaarheid van digitale media nog veel onzekerheden bestaan', zijn er toch voorbeelden van projecten waar voor digitalisering is gekozen. En daarmee bedoel ik natuurlijk niet zoiets als wat onze eigen KB in 1995 deed, toen *Honderd hoogtepunten van de KB* tegelijkertijd als coffee table book én op Internet verscheen: dat was vooral een visitekaartje, een publiekstrekker. In het kader van preservation management gaat het altijd om grootschaligheid.
Een - nog steeds bescheiden, maar sympatiek - voorbeeld geeft IJsland: een land met minder dan 250.000 inwoners en een navenant klein taalgebied, dat niettemin een groot cultureel erfgoed heeft te beheren: vele oud-germaanse sagen, voornamelijk in manuscriptvorm overgeleverd. Het behoud van deze collectie heeft tot op heden nooit bijzondere problemen opgeleverd, maar de toegankelijkheid voor een groot publiek was vanwege de ouderdom en uniciteit der documenten zeer beperkt. In het project SagaNet, begonnen in juli 1997, worden nu 565.000 bladzijden gedigitaliseerd en te zijner tijd via Internet openbaar gemaakt. De nationale bibliotheek is terecht trots op dit project, maar ook bezorgd: voor zo'n klein land is het een grote financiele inspanning, en de toegankelijkheid zal vrijwel zeker slechts van korte duur zijn, waarna migratie of reformatting weer een fortuin gaat kosten.[8]
Het tot nu toe grootste en langst lopende digitaliseringsproject werd gerealiseerd in het Archivo General de Indias (AGI) in Sevilla. Het AGI werd in 1781 opgericht - en is nog steeds gevestigd in een zestiende/zeventiende-eeuws gebouw - met als doel alle documenten te beheren die betrekking hebben op het Spaanse bewind in Amerika en op de Filipijnen. Het bevat ruim honderd miljoen bladzijden, wordt per dag gemiddeld door 50 onderzoekers bezocht en verstrekt jaarlijks 300.000 tot 400.000 kopieën op papier of microfilm. Om zowel behoud als toegankelijkheid te bevorderen werd in 1986 een digitaliseringsproject opgezet dat, niet toevallig, in 1992 officieel in gebruik werd gesteld, en ruim elf miljoen gedigitaliseerde bladzijden opleverde. Die elf miljoen vormen dus slechts 10% van het hele bezit, maar zijn op grond van gebruikersonderzoek zodanig gekozen dat thans dertig procent van alle raadplegingen elektronisch kan worden afgehandeld, wat dus inderdaad belangrijke winst oplevert voor zowel toegankelijkheid als behoud van de meest geraadpleegde documenten. Het systeem is intussen bijna zes jaar in gebruik. Digitalisering van documenten begon in 1989, en onvermijdelijk hebben er in de tussenliggende tien jaar diverse migraties plaatsgehad. Zo begon men met elkaar snel opvolgende (en allemaal erg dure) generaties WORM-disk, die nu zijn vervangen door intussen veel goedkoper geworden CD-ROM.
Op zijn zachtst gezegd opvallend is dat in het hele projectverslag (Gonzáles 1999) geen kritisch of bezorgd woord in verband met digitaliteit te lezen valt. Zo wordt anno 1999 zonder blikken of blozen meegedeeld dat backup-kopieën worden opgeslagen op magneetband, alsof in de tien jaar van dit project de onbetrouwbaarheid van dit medium niet met zekerheid is vastgesteld. Ik geef nog maar eens een citaat: in het rapport *Digitisation as a method of Preservation?* (Weber en Dörr 1997) schrijft de ECPA in het voorwoord: 'Yet, for preservation managers digitisation is in a way a wolf in sheep's clothing. How to deal, from a preservation point of view, with a medium that is notoriously unstable, for which 10 years is a long term'.

Papier > papier

Er is nog een andere mogelijkheid om een boek een tweede leven te geven. Ik had het zelf bedacht en ongetwijfeld zijn velen op het idee gekomen, maar de eerste die ik de gedachte op een internationaal congres[9] hardop hoorde uitspreken en toelichten was Bernhard Fabian, emeritus hoogleraar aan de Universiteit van Münster, en dus krijgt hij de eer te worden geciteerd (Fabian 1996, p. 35-36):

The current options are microreproduction and electronic storage - both highly problematic. The real danger, as I see it, is that these are regarded as the only options and that the replacement of the book by the microtext and the replacement of the microtext by the electronic text is more and more felt to be desirable and, by a further narrowing of the perspective, inevitable. (...) Reprinting is a well-established process and replaces a book by a book. In principle, reprinting is the best method available to counteract the decay of the printed tradition. And it is also the best means we have to ensure that texts to be read will be read and, above all, *can* be read.

De cursivering is van mij, SH, omdat ik wil benadrukken dat inderdaad alleen van een tekst gedrukt op zuurvrij papier, *en van geen enkel ander medium*, zeker is dat hij over 500 jaar nog gelezen kan worden.

Digitaal > digitaal

Zoals al eerder opgemerkt/geciteerd, wordt digitale opslag gekenmerkt door grote problemen, en wel van verschillende aard. Sommige daarvan, zoals juridische, zijn soms gebonden aan landsgrenzen (zie *Auteursrechtelijke aspecten van preservering van elektronische publicaties*, 1998), technische zijn universeel. Het belangrijkste daarvan is de snelle veroudering van hard- en software, waardoor veel materiaal nu al verloren is gegaan, en ander materiaal in het gunstigste geval nog wel behouden zou kunnen worden, maar straks niet meer toegankelijk is. Een eenvoudig voorbeeld daarvan is de echte floppy disk, het soepele schijfje van 5" inch, dat minder dan tien jaar geleden nog algemeen gangbaar was, en nu in geen enkele computer meer past. De oplossing in dit geval was: de gegevens tijdig verhuizen ofwel migreren van 5" inch floppy naar 3" inch diskette. Zo zijn er vele, en veel complexere, problemen te noemen, waarvoor tot voor kort migratie als de oplossing werd gezien. Waters en Garrett (1996, p. ii) daarover:

Migration is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display and otherwise use them in the face of constantly changing technology. The Task Force regards migration as an essential function of digital archives.

Het blijkt echter dat niet alleen de technologie voortdurend verandert, maar ook de heersende opvatting over hoe deze rampspoed het hoofd te bieden - wat ons er voorlopig toe verplicht de status en waarde van beide met groot scepsis tegemoet te treden. Slechts drie jaar na het CPA-rapport van Waters en Garrett verschijnt het ECPA-rapport *Avoiding Technological Quicksand* van Rothenberg (1999), waaruit ik citeer:

There is as yet no viable long-term startegy to ensure that digital information will be readable in the future. (...) Preserving digital documents may require substantial new investments, since the scope of this problem extends beyond the traditional libary domain, affecting such things as government records, environmental and scientific baseline data, documentation of toxic waste disposal, medical records, corporate data, and electronic-commerce transactions. (p. [vii])

Het probleem is intussen niet wezenlijk veranderd, hoogstens nog wat verder gepenetreerd, maar de oplossing luidt nu heel anders. Rothenberg stelt voor, het digitale document op te slaan in zijn oorspronkelijke logische vorm, d.w.z de verzameling bits die de tekst, beelden etc. vertegenwoordigen; deze logische vorm moeten we onderscheiden van de vele fysieke vormen die het document kan aannemen, zoals magneetband, diskette, enz. Het document in zijn logische vorm moet vervolgens worden opgeslagen met en onlosmakelijk van het computerprogramma waarmee het werd vervaardigd (het moederprogramma) en een gedetaileerde definitie van de computer waarvoor het moederprogramma werd ontworpen. Zodoende zal op een latere computer de oude hard- en software kunnen worden nagebootst; vandaar de voor deze procedure gebruikte term *emulatie*.

In zijn voorwoord tot Rothenbergs rapport stelt onze landgenoot Edo Dooijes[10] dat dit ongetwijfeld de juiste weg

is om tot een oplossing te komen van het digitale opslagprobleem, maar ook dat er nog vele vragen onbeantwoord zijn. Het zijn vragen die mijn technische kennis en begrip te boven gaan; Dooijes, als zuivere wetenschapper, ziet ze als een intrigerend en uitdagend onderwerp van verdere studie, zowel vanuit academisch als industrieel perspectief.

Los van de migratie/emulatie-discussie worden er al geruime tijd pogingen in het werk gesteld om digitale bestanden te archiveren, m.n. ook het WWW. In de eerste plaats moet je dan beslissen of je een *keuze* zult maken, of *alles* gaat archiveren; bedenk daarbij dat selecteren arbeidsintensief is en dus duur, en computeropslag steeds goedkoper wordt;
bedenk daarbij ook dat het WWW door alle links één grote kluwen vormt waaruit bijna geen geïsoleerd bestand valt los te weken; bedenk daarbij tenslotte dat je, althans theoretisch, op dit ogenblijk het hele Web kunt kopiëren en opslaan, maar dat er morgen weer een ander Web is. Ondanks deze immense problemen wordt er gewerkt aan en met concrete programma's. Zo is er in Australië PANDORA, een van die prachtige acronymen waar het informatievak zo rijk aan is ('Preserving and Accessing Networked DOcumentary Rescources in Australia'), en zo is er PreWeb, een internationaal samenwerkingsproject van de nationale bibliotheken van Zweden, Nederland en Australië (PreWeb = Preserving the WorldWideWeb).

Microfiche > digitaal
Aan het einde van de eerder in deze paragraaf geciteerde tekst 'Waarom microverfilming?' uit de Metamorfoze-brochure lazen we: 'Een microfilm die in het kader van Metamorfoze wordt gemaakt, gaat minstens 200 jaar mee en voldoet aan de kwaliteitseisen die noodzakelijk zijn voor optimale digitalisering in de toekomst'. Dit suggereert dat onderzoek verricht is naar een en ander, en dat is natuurlijk ook zo.

In het literatuurrapport *Digitalisering van microvormen* (Ligthart 1999) worden in verband met deze problematiek vier aspecten onderscheiden, die ik hier in hoofdzaak indicatief weergeef. 1. Bij de *productie* van microfilms en -fiches moet al rekening worden gehouden met de latere digitalisering; die stelt namelijk bijzondere eisen aan variabelen als polariteit, rollengte, resolutie en densiteit. 2. De *scannerkeuze* wordt in hoge mate bepaald door de vraag of er snel grote hoeveelheden gelijksoortige microfilms gescand moeten worden, of incidenteel kleine beetjes. Dit verschil bepaalt voor een groot deel de aan de scanner te stellen eisen. 3. Bij de *opslag* van digitale bestanden zijn factoren als compressie, formaat en opslagmedium van groot belang; de keuze is sterk afhankelijk van het gebruiksdoel. 4. Ook met betrekking tot de vele aspecten van *beschikbaarstelling* is het gebruiksdoel dikwijls beslissend. In elk geval moet bij het maken van beschrijvingen voor digitale versies onderscheid gemaakt worden tussen technische en bibliografische metadata. De technische metadata-elementen moeten in een "digitale header' worden opgenomen en toegankelijk zijn voor de gebruiker.

4 Maatregelen op macro-niveau

Om de met behoud & toegankelijkheid samenhangende problemen dichter bij een oplossing te brengen, is veel geld nodig en een groot maatschappelijk draagvlak. De direct met deze problemen geconfronteerde instellingen (bibliotheek, archief, museum) hebben dat geld niet, en het draagvlak is klein: het probleem en de omvang ervan zijn in brede kring juist onbekend of onderschat.
Aan dat laatste valt iets te doen door publiciteit (periodiekjes als *Metamorfoze nieuws*, paginagrote advertenties), door grootschalige actie van wetenschappers (zij zijn de belanghebbenden en vormen een gezaghebbende groep) en andere pressiemiddelen. Cruciaal hierbij is een gezamenlijk optreden van de drie instellingen die zijn belast met het behoud van ons cultureel en intellectueel erfgoed. Met andere woorden: nodig is convergentie van die instellingen, onder het motto 'United we stand, divided we fall' (Rugaas)[11]. Alleen zo'n gezamenlijk assertief optreden kan bij overheid en bedrijfsleven geld genereren.

5 Rol van de informatiemanager

In de inleiding stelde ik dat ik dit artikel schrijf vanuit het perspectief van de informatiemanager. Met die term duid ik iemand aan die de algehele verantwoordelijkheid draagt voor collectievorming, -beheer en -behoud; soms is er sprake van een specifiek op het behoud gerichte functie, die ik bij gebrek aan beter nu maar even *preservation manager* noem. Voor beiden geldt dat de hierboven gedeeltelijk beschreven *state of the art* zowel duidelijke conclusies als grote vraagtekens oplevert. Wie zijn oor op de juiste plaats te luisteren legt, kan vaststellen dat stemmen die eind twintigste eeuw uit de beroepspraktijk klinken bijna unaniem van mening zijn dat er met betrekking tot behoud & toegankelijkheid vooralsnog een tweesporen- of hybride beleid gevoerd moet worden, omdat geen enkele technologie zowel het een als het ander verzekert: toegankelijkheid bevorder je door digitalisering, behoud door microfiche of -film.
Veel moeilijker is het om tot besluitvorming te komen over massa-ontzuring. Diverse langjarige experimenten maken weinig enthousiasme los, en worden maar niet echt grootschalig. Daar staat tegenover dat de Library of Congress - toevallig (of juist niet?) wel de grootste bibliotheek ter wereld, met grote expertise en strenge normen - een meerjarig contract voor massa-ontzuring verlengde. We zullen dit met argusogen moeten blijven volgen.
Dat betekent overigens niet dat de informatiemanager de chemische processen die bepalend zijn voor de Battelle- en Bookkeeper-methode tot in details moet kunnen doorgronden: hij is geen chemicus of fysicus.
Hij moet wel weten wat de respectievelijke voor- en nadelen zijn, en vooral ook wat een en ander kost. Een informatiemanager is evenmin een ICT-deskundige en hoeft niet per se uit te kunnen leggen wat het verschil is tussen migratie en emulatie; hij moet wel beredeneerd tussen die twee kunnen kiezen. Een informatiemanager moet van veel markten thuis zijn en als hij op enig gebied kennis tekort komt, moet hij in- of externe deskundigen raadplegen. Het is zijn taak om op grond van die verzamelde kennis besluiten te nemen en te handelen.

Ik ben van mening dat het hoger beroepsonderwijs voor Informatiedienstverleners en -managers (IDM) zodanig moet zijn ingericht dat het toekomstige informatiemanagers aflevert die aan deze criteria voldoen. Dit lijkt misschien een overbodige opmerking, maar feit is dat hierin niet wordt voorzien door het kerncurriculum dat door de zes landelijke IDM-opleidingen is overeengekomen en dat 70% van de totale leerstof omvat. Het Instituut voor Media en Informatie Management aan de Hogeschool van Amsterdam beoogt het curriculum zodanig te herzien dat een en ander in augustus 2000 gerealiseerd zal zijn.

6 Duurzaamheid van technologieën

Vandaag de dag is digitaliteit zonder twijfel de meest besproken en tot de verbeelding sprekende technologie; in groot historisch perspectief gezien lijkt digitaliteit zelfs het vierde Hoofdstuk in de vijfduizendjarige geschiedenis van de documentaire informatie-overdracht te gaan vormen: na oraal, scribaal en imprimaal nu dus digitaal.
We zouden ons nu dan bevinden in de overgangsperiode imprimaal-digitaal, waarvan het meest verrassende is dat die nu al enkele decennia duurt, terwijl digitalteit nog lang niet is uit-ontwikkeld: het ziet er waarachtig naar uit dat hij in duur gaat wedijveren met de incunabeltijd, die de transitie vormde van scribaal naar imprimaal.
Behalve die duur zijn er trouwens meer overeenkomsten, en ook anderen hebben daar al op gewezen (Dewar 1998). De eerste vijftig jaar na Gutenbergs uitvinding golden er m.b.t. het boek ook geen algemene standaards, en was door het ontbreken van titelpagina en paginering beschrijving en ontsluiting van, respectievelijk verwijzing naar een boek niet mogelijk. Iets vergelijkbaars zien we nu bij digitale bestanden.
Hiermee heb ik nog geen antwoord gegeven op de vraag 'welke ontwikkelingen in het begin van het volgende millennium een voorname rol zullen spelen', en dat doe ik niet ook, want ik weet het niet.

7 Duurzaamheid van instituties

Over de levensduur van instituties, met name die van de bibliotheek, meen ik op grond van argumenten wel iets te kunnen zeggen. Er zijn nu wereldwijd duizenden bibliotheken die samen honderden miljoenen boeken beheren; vele van die boeken worden al eeuwen, vijfhonderd of duizend jaar lang zorgvuldig bewaard. In het derde millennium worden die boeken niet plotseling opgeruimd, en ze worden ook niet allemaal gedigitaliseerd: dat is duur en bovendien nergens voor nodig.
Internet of niet: de laatste decennia is de productie van boeken niet af- maar toegenomen. Van die hoeveelheid zal ongetwijfeld een deel in de papiermolen terechtkomen, maar een ander deel wordt toegevoegd aan bovengenoemde bibliotheken: die raken dus niet leger maar voller.

Het is dus geen tijd om bibliotheken af te breken, maar om nieuwe te bouwen. In 1998 betrokken enkele van de grootste Europese bibliotheken nieuwe gebouwen: de British Library in Londen (een gebouw van anderhalf miljard gulden), de Bibliothèque National in Parijs, de Deutsche Bibliothek in Frankfurt am Main (met een huidig bezit van negen miljoen boeken en magazijnruimte voor nog eens negen miljoen *nieuwe*).
In 1999 werd de nieuwe Koninklijke Bibliotheek in Kopenhagen geopend. Kenmerk van al deze nieuwe gebouwen is, dat ze zijn geconcipieerd als attractieve en relatief laagdrempelige kennis- en cultuur centra. De Deense KB is een gebouw volgens een nieuw concept: ruime entree, grote toegankelijkheid, concertzaal en andere culturele activiteiten (Nielsen 1999); in Frankfurt is het gebruik met 60% toegenomen. (Stephan 1999). Deze investeringen zijn niet het gevolg van conservatisme en ook niet van de waan van de dag, maar bedoel om een cultureel instituut voort te zetten dat niet alleen millennia oud is, *maar ook vandaag de dag nog uiterst adequaat.*
Ook in het jaar 3000 zullen er dus nog bibliotheken zijn.

Literatuur en andere informatie

Twee afkortingen: CPA Commission on Preservation and Access
ECPA European Commission on Preservation and Access

Voor wie op de hoogte wil raken of blijven van de onderhavige problematiek is de ECPA een onmisbare bron van informatie.
Postadres: Postbus 19121, 1000 GC Amsterdam
Bezoekadres: Trippenhuis, Kloveniersburgwal 29, Amsterdam
Telefoon: 020 551 08 39
Website: http://www.knaw.nl/ecpa/

Auteursrechtelijke aspecten van preservering van elektronische publicaties (1998), Amsterdam: Instituut voor Informatierecht. IviR Rapporten - 7

Bakker, Mart, & Alfred Schmits (1997), 'Het digitale erfgoed', in: Pieter Wisse, *Het boek van bewaring : Selectieve duurzaamheid in de informatiemaatschappij,* Alphen aan den Rijn: Samsom Bedrijfsinformatie, p.91-109

Bellinger, Meg (1999), 'Preservation Resources looks to future through past', in: *OCLC Newsletter* No. 238 (March/April 1999), p.24-25

Conserveren in de toekomst (1998), themanummer van *Informatie Professional* [jrg. 2 (1998), nr. 10], voorbereid door Gerda Huisman, Marianne Pothoven en Jenny Mateboer. De blz. 19-43 bevatten de volgende bijdragen: Pierre Pesch, 'Conservering in de bibliotheek', Yola de Lusenet, 'Conserveren is vooruitzien', Freek van Eijk, Geert Maas en Jenny Mateboer, 'Bouwen voor de eeuwigheid', Mariska Herweijer, 'Digitale dilemma's', Trudi Noordermeer, 'Het reguleren van de tand des tijds' Chris Groeneveld, 'Auteursrecht versus behoud van elektronische publicaties'

Dewar, James A. (1998) 'The printing press and the networked computer: Parallels that may illuminate the future', in: *Logos,* jrg. 9, nr. 4, p. 187-194

Dooijes, Edo H. (1995), 'De computer is rijp voor het museum', in: *Collecties : Jaarboek 1995 van de Universiteit van Amsterdam,* Amsterdam: Vossiuspers AUP, p. 50-55

Fabian, Bernhard (1996) 'Preservation - A Personal View', in: Yola de Lusenet (ed.), *Choosing to Preserve : Towards a cooperative strategy for long-term access to the intellectual heritage. Papers of the international conference in Leipzig, March 29-30, 1996.* Amsterdam: ECPA, p. 17-37

Feather, John (1991), *Preservation and the management of library collections.* London: The Library Association

Feather, John, Graham Matthews and Paul Eden (1996) , *Preservation Management : Policies and Practices in British Libraries,* Aldershot: Gower

González, Pedro (1999), *Computerization of the Archivo General de Indias : Strategies and Results.* Washington: Council on Library and Information Recources / Amsterdam: ECPA

Havermans, J.B.G.A.(1996), *Haalbaarheidsstudie Bookkeeper Proces. Deel 1: Literatuurstudie Effecten Alkali op de Versnelde Veroudering van Papier.* Delft: TNO Paper and Board Research.

Havermans, J.B.G.A. en J.P. van Deventer (1996),
Haalbaarheidsstudie Bookkeeper Proces. Deel 2: Praktijkonderzoek, Procescontrole en Discussies. Delft: TNO Paper en Board Research

Havermans, J.B.G.A (1997).,
Deacidification using the Bookkeeper process. TNO-report BU2.97/009010-1/JH. Delft: TNO, 9 December 1997. 16 p.

Horsman, Peter (1998),
Digitaal Archiveren : Het Recordkeeping Systeem als kader voor het beheer van digitale archiefbescheiden. Den Haag

Hubregtse, Sjaak (1997),
• Preservation & Access: State of the Art and Future', in: *New Book Economy: proceedings of the 5th international BOBCATSSS symposium, Budapest, January 1997,* Amsterdam: Hogeschool van Amsterdam, p. 115-123

Küller, Elle (1997),
• *Verantwoording verzekerd: Het programma digitale duurzaamheid'*, in: Pieter Wisse, *Het boek van bewaring : Selectieve duurzaamheid in de informatiemaatschappij,* Alphen aan den Rijn: Samsom Bedrijfsinformatie, p.73-90

Ligthart, Anita (1999),
Digitalisering van microvormen : Onderzoek naar de verschillende aspecten van de conversie van microvormen naar een digitale omgeving: een omgevingsverkenning, Den Haag: Koninklijke Bibliotheek, 1999

Lusenet, Yola de (ed.) (1997),
Choosing to Preserve : Towards a cooperative strategy for long-term access to the intellectual heritage. Papers of the international conference in Leipzig, March 29-30, 1996. Amsterdam: ECPA

Memory of the World : General Guidelines to safeguard documentary heritage (1995)
Prepared for UNESCO on behalf of IFLA by Stephen Foster, Jan Lyall, Duncan Marshall and Roslyn Russel. Paris: UNESCO

Metamorfoze nieuws (juni 1997...)
[nieuwsbrief Metamorfoze Programma, Den Haag: Koninklijke Bibliotheek]

Nielsen, Erland Kolding (1999)
'A Concept for Collection Management, Old and New Media', in: *Preservation and Access : Nordic Conference on Preservation and access held at the Royal Library in Stockholm October 5-6, 1998 arranged by the National Libraries of Denmark and Sweden,* p. 23-28. Kungl. Biblioteket Rapport Nr 25, 1999-10-01

Ontzuren op een behoedzame wijze (1999)
[brochure van Archimascon B.V. / Bookkeeper]

Porck, Henk J. (1996),
Mass Deacidification : An Update of Possibilities and Limitations, Amsterdam: ECPA / Washington, CPA,

Preservation Management : Between policy and practice (1999),
[programma met voorlopige abstracts van congres, gehouden in de Koninklijke Bibliotheek op 19-21 april 1999]

Principles for the Care and Handling of Library Material (1998),
Compiled and edited by Edward P. Adcock with the assistance of Marie -Thérèse Varlamoff and Virginie Kremp. Paris: IFLA-PAC

Rothenberg, Jeff (1999),
Avoiding Technological Quicksand : Finding a Viable Technical Foundation for Digital Preservation, Washington: Council on Library and Information Resources / Amsterdam: ECPA.

Rütimann, Hans (1994),
'Saving the memory of humanity: A crisis in the world's libraries', in: *Logos*, jrg, 5, nr. 4. P.166-171

Saving the Written World: Mass Deacidification at the Library of Congress (1999), Washington: Preservation Directorate, Library of Congress [vouwblad]

Schouten, Dennis, en Ingeborg Verheul (1998),
Metamorfoze : Vechten tegen het papierverval (Actuele Onderwerpen nr. 2657). Lelystad: AO BV, 30 oktober 1998

Schwartz, Werner (1996),
European Register of Microform Masters(EROMM) : Supporting International Cooperation,
Amsterdam: ECPA

Stephan, Werner (1999)
'Die Deutsche Bibliothek: House of Books and Electronic Archives. Changes in Services as a Response to New Technology and New User Demands', in: *Preservation and Access : Nordic Conference on Preservation and Access held at the Royal Library in Stockholm October 5-6, 1998 arranged by the National Libraries of Denmark and Sweden*, p. 72-78. Kungl. Biblioteket Rapport Nr 25, 1999-10-01

Vuijst, J. De, en J.S. Mackenzie Owen (1997),
'Digitale houdbaarheid', in: *Jaarboek voor Nederlandse boekgeschiedenis 4 (1997),*
p. 191-206

Waters, Donald, en John Garret (1996),
Preserving Digital Information : Report of the Task Force on Archiving of Digital Information,
Washington:CPA

Weber, Hartmut, en Marianne Dörr (1997),
Digitisation as a Method of Preservation : Final report of a working group of the deutsche Forschungsgemeinschaft, Amsterdam: ECPA / Washington: CPA

Weten geweten gewist : Bedreigde wetenschappelijke collecties in archieven en bibliotheken (1997)
Tekst en samenstelling: Gabriëlle Beentjes, Mariska Herweijer, Yola de Lusenet, Karin Scheper en Paula Witkamp, Amsterdam: ECPA

Noten

1 De inhoud van par. 1 was eerder, in een uitvoeriger versie, onderdeel van mijn bijdrage aan het vijfde BOBCATSSS-symposium in Boedapest (Hubregtse 1997).

2 Zie m.n. het onder redactie van het Instituut verschenen boek *Toeval of Noodzaak? : Geschiedenis van de overheidsbemoeienis met de informatievoorziening,* Amsterdam: Otto Cramwinckel, 1995.

3 'Intellectual Property vs The Right to Knowledge', 8th international BOBCATSSS Symposium on Library and Information Science, 24 t/m 26 jan. 2000.

4 Voorbeelden van deze en vele andere rampen zijn te zien in de film *Into the Future: On the Preservation of Knowledge in the Electronic Age* (1997) van Terry Sanders, uitgebracht door de Council on Library and Information Resources i.s.m. de Commission on Preservation and Access en de American Council of Learned Societies.

5 Begeleidende tekst, opgesteld door de Council on Library and Information Resources, bij de in noot 4 genoemde film.

6 'Bookkeeper' is een gepatenteerd systeem van het Amerikaanse bedrijf Preservation Technologies, Inc. Een licensie is verstrekt aan de Nederlandse firma Archimascon bv, die in Heerhugowaard een kleine opstelling in gebruik heeft.

7 Gesprek dd 20.04.1999 met James E. Burd, directeur van Preservation Technology.

8 Thorstein Hallgrímsson , directeur techniek van IJslands Nationale Bibliotheek, op de 'Nordic Conference on Preservation and Access', Stockholm 5-6 oktober 1998.

9 Het eerste ECPA-congres, 'Choosing to Preserve', Leipzig 29-30 maart 1996.

10 Edo H.Dooijes is docent aan de vakgroep computersystemen, en conservator van het computermuseum, beide aan de Universiteit van Amsterdam. Zie ook Dooijes 1995.

11 Op 13-14 aug. 1998 vond als 'satelite event' van het grote IFLA-congres in Amsterdam het seminar 'Convergence in the digital age' plaats. Daar en elders (Stockholm 1998, zie noot 8, Den Haag 'Preservation Management' 1999) is Bendik Rugaas, bibliothecaris van de KB Oslo, de onvermoeibare pleitbezorger van convergentie in deze zin.

Modular scenarios in the electronic age

1) Meta-Information

1.1) Title: Modular scenarios in the electronic age

1.2) Authors: Joost G. Kircz and Fédérique Harmsze

1.3) Address: Van der Waals-Zeeman Instituut, Universiteit van Amsterdam
Valckenierstraat 65-67, 1018 XE Amsterdam, The Netherlands

1.4) Internet Address: {Kircz, Harmsze}@ wins.uva.nl
www.wins.uva.nl/projects/commphys

1.5) Publication: Conferentie Informatiewetenschap 2000

2) Positioning

The scientific community is confronted with a manifest information overload. Due to the great advances in electronic communication, the fear of a massive information infarct only becomes more real. However, as all new technologies strongly influence the way information is created, organised and presented, we can try to rescue the situation by taking the intrinsic properties of the new electronic technology as a starting point for a novel communication framework.
In an electronic environment, information is stored according to schemes independent of the consecutive space or time order of the original information and is therefore randomly accessible. In this contribution, we develop the consequences of these essential characteristics of distributed storage and random retrieval for information handling in a bottom-up approach. We first briefly survey the coming into being of the present paper-based situation. We then proceed with the conjecture that the future of electronic scientific communication will be characterised by modularity. Our aim is to design such a new, modular, framework, in which documents are presented as a collection of coherent entities. We discuss at length the idea that, in an electronic environment, links between modules become information objects on an equal footing with modules themselves, with the concomitant need to design representation classes. We then present the outline of such a framework. In this presentation, we emphasise the more general aspects than those presented in an in-depth study in physics that was recently published. The results presented here will hopefully serve as ingredients for applications in different fields in natural science or medicine. Based on these general notions, concrete designs, with a strong accent on authoring tools, can be made.

2.1) Situation: What is it all about?
Technology has always both helped and hindered human communication. In person-to-person contact, speech and gesture establish a unified framework. In oral society, stanza, rhythm and rhyme were important mnemonical techniques for structuring information [Ong82]. Information was communicated according to rules governed by the need that information must be memorised by humans. With the transition to literacy, the communication process changed fundamentally. Written text and illustrations liberated the human brain from the task of memorizing lengthy texts. It established new ways of structuring information closer to the intrinsic character of the information itself. Thus, e.g., long lists of data no longer demanded all kinds of memory aids, but were now printed in their full, dull and monotonous nature. The one-dimensional time axis of speech is now mapped onto the two-dimensional plane of the written text. The unique qualities of paper enabled the invention of elaborate 2D-graphical representation of data, which introduced its own pictorial rhetoric [Tuf83, Tuf90]. Writing developed its own style, away from pure speech [Hav86, Mcl62].

However, the ability of the reader to jump haphazardly through a text does not imply that in writing, communication stops following the old oral concept of the linear ordering of a trail of thought. A well-structured argument still follows sequential steps. Style manuals of all kinds insist on the logical sequential order of the unrolling story. The author writes for the unknown reader and therefore creates a path that can be followed by everyone. For a discussion on the different roles of authors and readers, see [Kir91]. The reader, be (s)he an informed reader or an ignoramus in the field, hardly starts reading at line one to follow diligently the entire text. Before the decision is taken to read a text, it is scanned and browsed through. People then decide to skip the text, to read parts of it or to read it in full. In contrast to listening, reading allows for non-linear and haphazard consumption. Associative and lateral thinking is much easier using the two-dimensional plane of paper, than the forced-march route of the one-dimensional time-axis of spoken language. Due to paper, the selective consumption of information took off, as paper allows for non-sequential reading and skipping large parts.

Interestingly the need for certified information, necessary for maintaining the integrity of published texts, induced very strict rules for the publishing of non-fiction and in particular scientific texts. The role of technology, in particular the printing press, in shaping texts and distributing knowledge and information has been extensively studied by [Eis79, Eis83]. The role of the necessary transport material, i.e., paper, in the same way that the development of the car demands the development of roads, is the starting point of [Feb76]. The role of multiplication as the prime mover for the dissemination of all kinds of semi- and even non-scientific information as motor for the transcending of pre-scientific thinking to modern scientific practice is emphasized by [Eam94].

As a form of research communication, the scientific paper is particularly interesting, as it demands a high level of clarity, integrity and standardisation. In the following, we will narrow down our discussion to this type of work, although many observations and remarks may also hold for other types of work, especially in the educational realm. Much has been written on the development of the scientific article as carrier for scientific information transfer [Mea98 and references therein]. In an earlier article [Kir98], we argued that the linear form of the scientific article is a typical consequence of print on paper.

With the emergence of electronic media, the two-dimensionality of paper is enhanced by new parameters that allow the free rearrangement of every piece of information. This ability for nonlinear reading is rooted in the way electronic information is stored. Independent of the intended order of the intellectual content, the bit-representation is essentially nonlinear and can even be distributed among geographically disconnected memories. This new fact allows the reader not only to have an individual reading path through the work, but also allows for the addition of comments, extra -secondary- information to, and the expansion of every single part of a particular work.
Another important advantage of modularity is that for the searching reader, keywords are put into context. In traditional information retrieval a query composed of a combination of keywords results in the retrieval of a series of more-or-less relevant documents. In a modular environment the name of the module adds context to the keyword. Now explicit queries can be envisioned like "Keyword" AND "Module X (e.g., Results)" NOT "Module Y (e.g. Introduction)".

2.2) Central problem: A new granularity and its structure

An important precondition for the exploitation of these dynamic capabilities is that we introduce a well-defined *granularity* of information. The design of such a new structure has to be based on general accepted ideas about the consistency and integrity of scientific communications, as well as on the analysis of standard scientific papers. On every level of granularity, there is a need for *surveyability*. Especially in an electronic modular environment, where authors can add their own modules with new work or can add comments to old works, clear traffic signs are needed. The purpose of this paper is to discuss the new opportunities and to present a general framework for modules and their relations that enhance the research communication in an electronic environment.

3.1) Methods: General ideas on modules

As indicated above, the typical 'grain size' in a print-on-paper environment is the scientific paper as we know it. A consequence of the electronic environment is the introduction of a new type of granularity of information. A new granularity that implies new ways of ordering: ordering that mimics the old trusted methods, and also a structuring that exploits the new capabilities. As with print-on-paper, we face, on the one hand, the aspects of reproduction, which is the representation of an object made using a certain technique into another one, e.g., the photograph of an oil painting. Reproduction is specifically the case if some form aspects of the object are essential. For instance, the text of a speech or song will always be intended to be read aloud and in a fixed order, even if it is printed on paper or stored in electronic form. On the other hand, in the electronic realm, we also encounter novel ways to structure information and knowledge that are unique to the medium, like elaborate data tables that are became possible on paper and CAD/CAM simulations that are unique to electronic media.

For scientific research papers, we can distinguish modules of information that represent self-contained, though related, types of information. Writing a modular article does not mean a simple "cutting-up" of an existing linear text, but imply a different grouping of the same information. As the reasoning of the author needs a sufficient length of text to present the message in an intelligible way, modules of information have no intrinsic length (or in the case of pictorial information, no surface) limits. The essence of the idea of modularity is that modules comprise cognitive units. We reported a first approach to such a modular structure in [Har96], a in-depth analysis of the concept of modularity is given in [Kir98].

Thus, because of the fact that information in an electronic environment can be accessed from many different directions, re-used piecewise and augmented on demand, a novel structure is needed. Hence, a new class of descriptors is needed. Such a novel system must tally with the standard requirements for scientific communication with regard to clarity, quantity, quality and relevance. In the *Communication in Physics Project* [Har99, Har00], an elaborate attempt has been made for such a new structure in the field of molecular physics. We analysed a coherent corpus of physics articles with the aim of creation of a new model for electronic articles in experimental sciences based on selective reading and multiple use of information. In this project, we propose a scheme for writing scientific papers in a modular form, including instructions for authors. A full description of this model is given in the PhD thesis of Frédérique Harmsze [Har00]. Below, we rely on this work, but try to describe our ideas on a more generic level.

Following [Har00], we define a *module* as a uniquely characterised, self-contained representation of a conceptual information unit aimed at communicating that information. Not its length, but the coherence and completeness of the information it contains make it a module. This definition leaves open that modules are textual or, e.g., pictorial. Modules can be located, retrieved and consulted separately as well as in combination with related modules.
Elementary modules can be assembled into higher-level, *complex modules*. We define a complex module as a module that consists of a coherent collection of (elementary or complex) modules and the links between them. Using a metaphor, elementary modules are 'atomic' entities that can be bound into a 'molecular' entity: a complex module.
We distinguish two types of complex modules: *compound modules* and *cluster modules*. In a *compound module*, related (albeit possibly dissimilar) modules are aggregated to form a new module on a higher level. An example of an aggregated module is the module 'Experimental methods' that can be composed of lower-level modules representing the various components of a measuring device. In our physics corpus, we encounter, e.g., a molecular beam apparatus that has relatively independent components like: one or more sources of a particle beam, an interaction chamber, and a detector.
The central idea of a *cluster module* is the generalisation of specific concepts that its constituent modules focus on. An example of a cluster module is the module 'Raw data', composed of various elementary modules reporting the results of the same general type of measurements involving different molecules.
A direct consequence of splitting up the different kinds of information is that in complex modules, we need some extra text, summarising the essence of the composing modules in so-called module summaries. These module summaries play also an important role in the creation of an abstract for the entire modular article. In a modular environment, the abstract is particularly important, being a concise expression of the coherence of the discourse.

This aspect of modularity is the subject of a related, ongoing, project [Tol99]. Summaries will be tagged separately and hence can be skipped by the reader on demand.

3.2) Methods: General ideas on relations and links

As mentioned in the Section 2.2 (Central Problem), on every level of granularity of information there is a need for surveyability. Especially in an electronic modular environment, where authors can add their own modules with new work or can add comments to old works, clear traffic signs and road maps are needed. Traditionally, we have meta-data as the representation or identifier of a piece of information. Meta-data describe the type of information contained in a unit (a paper, a painting, a module). Various classes of meta-data exist and preferably form a complete and non-overlapping system of coordinates, e.g.: bibliographic information, domain-specific keywords, etc. Information Retrieval (IR) is the science and art of manipulation of these meta-data in order to meet a more or less well-defined query (information need) of a reader. In a distributed environment, we need new kinds of meta-data: meta-data that express the relations between different modules. As far as it indicates the organisational structure of the related modules, this is, in essence, an extension of the traditional bibliographic meta-data. Just as a paper is part of a journal section, a journal section part of a journal, and a journal part of a publication programme, will a module be part of a modular article and a modular article part of a larger data base of modular publications.

The novelty of a modular environment is that we define modules as conceptually separate entities within an organised structure. Hence, the conceptual relations between modules are primary to the organisational ones. This means that, next to organisational relations, a series of discourse relations based on the communicative function of the message, as well as the content, have to be introduced. The expression of such relations in an electronic environment is given through explicitly-labelled links. We define a link as a uniquely-characterised, explicit, directed connection, between entire modules or segments of modules, that represents one or more different kinds of relations. A link then represents not only the organisational relations between two information units, but also harbours information on the why and wherefore. As a result, the various kinds of relations expressed in hyper-links lead to special classes of meta-data describing the attributes of that link. An immediate advantage of this approach is that from an IR point of view, we can keep all the machinery and just add these new types of meta-data to our system.

Thus, in a modular electronic information system, hyper-links are objects that represent particular relations and do not just represent the fact that something is somehow related to something else. A hyper-link becomes is endowed with information that expresses the relations it represents. One link can represent many relations, such as: the target belongs to the same work, contains a more elaborate account and a different presentation of the same data. A link might be created as part of a scientific report, but can also be created later by somebody who links already-existing material with new material. These new links to old material can represent important steps in the scientific discourse. Therefore, every link also has to carry the bibliographic meta-data of its originator. This way, a hyper-link becomes a new type of information object carrying particular meta-data, representing the coherence of the distributed information. With an acknowledgement to particle physics, one could call hyper-links *messenger-objects*. This brings us to an important consequence of modularity: *in an electronic modular environment links are objects on the same footing as modules.*

From the beginning of Hypertext, authors are aware of the need to structure the ever- increasing number of linked objects. A serious problem with all attempts to attack this intrinsic complication was the impossibility of proper implementation and therefore experimentation with linking schemes. The latest developments within the World Wide Web Consortium projects suggest serious breakthroughs. First we have the development of the Resource Descriptor Framework as a meta-data framework in which it becomes possible to add attributes and their values to any web resource [RDF99]. Secondly we have the development of XLINK, of the W3C XML linking Working Group [XML99]. These developments give confidence to the discussion on link systems becoming an essential ingredient in the design of future communication systems. For most schemes discussed in the literature, it is clear that at least two classes of link description are needed, one describing the organisational relations and another dealing with meaning and understanding. All authors call upon linguistic tools, however, without a systematic analytic approach. For that reason, most of the suggestions remain highly ad hoc. In many cases, an attempt is made to provide a complete set of all possible link types. An early, and oft-quoted, taxonomy can be found in Trigg [Tri83]. Here 70-odd types of links are classified in two categories. The first category contains *normal links* that "serve to connect nodes making up a scientific work, as well as to connect nodes living in separate works", whilst "*commentary links* connect statements about a node to the node in question".

The problem with this taxonomy is that it is of a pure phenomenological kind without any attempt to structure links according to some deeper understanding of speech communication research. In an attempt to structure hypertext for Classics and Religious studies, DeRose presents a more elaborate link taxonomy, split into the two main categories: *extensional links* and *intensional links* [Der89]. The *extensional links* are further split into *relational links* and *inclusive links*. Relational links consist of, first, *associative links* that connect arbitrary pieces of text and can be considered to follow a discourse, for which many types named by Trigg could be useful. Secondly, relational links contain *annotational links* as referential links that represent connections from portions of a text to information about the text. In the inclusive link realm, DeRose lists *sequentional links* and *taxonomic links* that associate lists of properties with particular document elements. In the second main branch of intentional links, DeRose lists link types that follow strictly from the structure and content of the document. Though very ingenious, the scheme lacks sufficient transparency for easy adoption and is insufficiently fine-grained, particular in the case of relational links.
In the work of Baron et al. [Bar96] tests were executed on a limited corpus of the OCLC Cataloguing User's guide, augmented with labelled links according two three classes: *semantic* (similar, contrast, part of), *rhetorical* (definition, explanation, continuation, illustration and summary) and *pragmatic* links (warning, prerequisite, usage, example). Their tests show that more effective searching becomes possible. A final example of an ad hoc approach is the usage by Rutledge et al. [Rut00] of the extensive list of rhetorical indicators by Mann et al. [Man88, Man89], to assist the structuring of multimedia presentations. Furthermore, here, a whole list of possible relations between information objects is used.

In an earlier paper [Kir91], we suggested a structured taxonomy of the lines of reasoning in scientific texts as a complement to the semantic networks of keywords. In this early work, a clear distinction between coherent content and relations between this content is not yet made.
The main argument was that a structural taxonomy was badly needed and preferred above a mere taxonomy of rhetorical indicators.

In the *Communication in Physics Project*, we are collaborating with the University of Amsterdam, Department of Speech Communication, Argumentation theory and Rhetoric, to try to find a systematic way of designing a linking taxonomy based on a linguistic approach. An important aspect is that the analysis will be based, not only on formal structural relations, but also on the pragmatic approach of speech communication. Here, the actual usage of language is the starting point, in the framework of a critical assessment of an argumentation that legitimates the step from premises to standpoints. Scientific communications are texts in which an author wants to convey a problem, solution or opinion within the context of a broader scientific quest. For that reason, large parts of scientific papers can be considered as argumentative texts. In a working electronic environment, we need a systematic structure in which the relations between different parts of the unravelling story have a clear meaning. Furthermore, because we see links as important information objects in themselves, we need to limit their number to create an efficient and effective authoring environment. The results presented in this contribution give the general outline of a structure of classes of modules and classes of their relations, which can serve as a starting point for more detailed domain-specific applications.

4) Results: Descriptor classes in a modular framework

As information in an electronic environment has essential nonlinear characteristics, we have to develop a framework in which the unique characteristics of particular information modules are well-defined in order to let different kinds of modules be handled according to their own, unique, aspects. In order to build new information systems, we have to analyse the various ways in which that information can be represented and identified. Below, we present an approach, in which we distinguish between different classes of representation and identification. As every kind of module can have features belonging to different classes, classes have to be of a different and mutually exclusive nature. These class-dependent features can then be structured in controlled keyword lists or thesauri per class, and can be considered as meta-data. It is important that the classes do not overlap in meaning; in mathematical terms, one would say that the classes represent orthogonal representations, and in every representation, with specific parameters. In the analysis, an important aim is also to ensure that the collection of classes is exhaustive and economical. Next to classes identifying the content and form of an information item, we also have to identify the classes describing the mutual relations between these items, as argued above. A complication that will not be addressed further in this paper, is that metadata can be attached to a module or to a

particular part of a module, for instance a word, a picture element, or a specific number.

4.1) Module descriptors

In order to be able to determine what is 'similar information' to be grouped together and represented in a self-contained module and, subsequently, to determine how to tag the resulting module, we need an unambiguous typology of scientific information. The analysis of Harmsze [Har00] and the subsequent testing of the model by rewriting physics articles shows that the following types are sufficient for the domain analysed and by proxy for most experimental sciences. Obviously in different fields the exact structure and definition of the modules can vary.

1-Bibliographic information

It goes without saying that our first class of identification is the traditional bibliographic data. As, in a multi-media environment, the presentation form of the same object can change, I prefer to confine the class of bibliographic data to data strictly concerning the author, the publisher, title, dates, etc. All references to the appearance of the work should be in a separate class.

2-Presentation forms

It is obvious then, that the next division in types of information will be according to appearance: text, figure, graph, photo, holograph, sound, simulation, etc., etc. The accompanied meta-data are items like presentation type, file length, colour scheme, frequency range, language type, resolution, etc. This division in "physical" appearance of information objects does not tell us anything about what is communicated. In the *Communication in Physics* project, we concentrate on the information, regardless of its presentation form; hence this class was not touched upon.

3-Domain-dependent keywords

On the same level, we can introduce the classical domain-specific keyword and classification code systems. Both traditional classes of identifiers naturally keep their relevance in an electronic environment.

We introduced in the project two new classes of typology: a characterisation by the range of information and a characterisation by its conceptual function, i.e., by the role the information plays in the scientific problem-solving process.

4-Range-based information

By characterising information by its range, so-called microscopic, mesoscopic and macroscopic modules can be introduced.

4.1-A microscopic module represents information that belongs only to one particular article, e.g., information concerning the specific problem addressed in that article.

4.2-A mesoscopic module functions at the level of an entire research project; it is created for multiple use in several articles issued from the same project. For example, information about the experimental set-up that has been used in a series of experiments can be presented in a mesoscopic module and connected to several articles reporting experimental results. The same holds for general theoretical approaches or computational algorithms used in a series of investigations.

4.3-A macroscopic module represents information that transcends the level of the research project; this type of firmly established information is given, e.g., in books, lecture notes, etc.

5-Conceptual function

The main characterisation in a modular structure is the identification of information by its conceptual function indicating the different steps in the research process. Modules are self-contained, therefore, every module represents only one well-defined aspect of the discourse in the article. Of course, this self-containedness does not mean that one module is usually sufficient to understand the whole work. Modularity enables the reader to immediately zoom in to those aspects he/she is interested in. If so wanted, the whole work, i.e., all the related modules and, if needed, the necessary auxiliary information presented in meso- and macromodules, can be retrieved and read as if it were a traditional article.

As modules are meant to be coherent bodies of content, it makes sense to have only a limited number of members in this class. Our starting point is the prototypical section structure of scientific papers in experimental sciences: Introduction, Methods, Results, Discussion and Conclusions. This sequence represents the normal flow of a scientific narrative, but the way it is used in practice presupposes that the article will, indeed, be read sequentially from beginning to end. Based on this prototypical structure we have defined the following, more systematic, modules.

5.1- Positioning is a complex module consisting of two modules.
5.1.1- The module *Situation*, describing the embedding of the work, and
5.1.2- the module *Central Problem*, stating the why of the work in question.
In this complex module, all the information the reader needs to know about the background of the problem in question and the particular aspects dealt within the article, is grouped together. It is immediately clear that the module Situation, which reviews the embedding of the work, can be to a large extent replaced by a pointer, linking the work in question to a description elsewhere. Such an introductory text is a typical kind of meso-information. This way, the enormous redundancy of information presented in introductions of articles can be avoided. Obviously the module Central Problem is an essential module, as this module provides the intentions of the author, given the context. For an informed reader, this module can play a decisive role in the decision to drop the article or to continue reading.

5.2- Methods is a complex module that can be built up from separate modules representing:
5.2.1- the *theoretical*, *5.2.2- experimental*, and/or *5.2.3- numerical* methods employed. If an article is one of a series, a substantial part of the information about the methods can be represented in mesoscopic modules for multiple use; e.g., in a pure experimental article using a standard instrument and employing a standard theory, both the *Experimental Method* and the *Theoretical Method* can be described elsewhere. In other fields like, e.g., organic chemistry or pharmacology, different types of methods can be defined, e.g., *Preparatory Methods* or *Treatment Protocols*, respectively.

5.3- The complex module *Results* allows readers to inspect the results without further need to read the whole article.
5.3.1- One of its two constituents is the module *Raw Data*. In printed articles, these data are hardly ever published, as that would require too much space. In an electronic environment these data can become directly available to the reader. By accessing the data in this way, the reader is able to use the data without the preferred interpretation of the originator.
This enables the reader to merge his/her own data directly with the presented data for comparison and analysis. It also allows different people to apply different methods for data reduction to the same data.
5.3.2- The second constituent of the module *Results* is the module *Treated Results*. Here the raw data are handled according to the author's choice for data reduction and further treatment. The module *Treated Results* presents the smoothed data in the usual form in figures and tables, as we are familiar with from traditional journals.

5.4- The module *Interpretation* contains the core of the scientific reasoning in the article. Here, the author interprets the experimental results in the light of a theoretical model, e.g., by comparing them with theoretical results and experimental results obtained by others. An important observation in our analysis is that it is this module that maintains most of the characteristics of a classical paper. One can argue that our procedure in fact strips the traditional article from those components that can be presented as self-contained information or data entities. These modules do contain regular text and, if needed, reasoning and arguments.

The remaining core, the real scientific reasoning, argumentation and conjectures, remains an essay-like text. It is this part, in fact representing complex knowledge rather than quantitative information, that is the most difficult one to deal with.

5.5- Within the complex module *Outcome*, we distinguish two modules:
5.5.1-Findings, a compulsory module in which the author tries to answer the central questions stated in the module *Central Problem.*
5.5.2- We also have an optional module *Leads to Further Research*, in which ideas and suggestions for new work are expressed. It is clear that the module *Findings* represents the final results of the work, which together with the module *Treated Results* allows a reader to learn about what happened without the how and why.

6- Within the model developed in the *Communication in Physics Project*, we collect the earlier-mentioned traditional classes (bibliography, presentation form and keywords) as well as an abstract, a map of contents, list of references per module, and the acknowledgement in a separate module *Meta-Information.*

4.2) Link descriptors

If we want a crisp definition for classes of relations, it makes sense to differentiate between two main classes: organisational relations, dealing with the pure structural aspects, and scientific discourse relations, dealing with the content and pragmatic relations such as arguments.

1- Organisational relations

With respect to the class of organisational relations, we suggest defining, as a minimum, the following types of relations.

1.1- Hierarchical relations are asymmetric relations between an objects and there constituents, such as between complex modules and the elementary modules they contain.

1.2- Proximity-based relations reflex how close linked modules are. For instance between elementary modules that are part of the same complex module, article or larger collection of modules. An example can be links between a certain kind of measurement on different species reported in modules for each specie and collected in one complex module. Another example are photographs of the same animals in a different environment collected in one complex module.

1.3-Range-based relations exist between micro-, meso- and macromodules of the same or different communications.

1.4- Path relations that allow for the construction of different reading paths, e.g.: a first cursory reading and a second in-depth reading path. Also, with the help different discourse relations (see below), such different paths can be set out. In [Har00], a distinction is made between a so-called *sequential path*, connecting all modules available and an *essay path* mimicking the reading of a linear document.

1.5- Representational relations are between different representations of the same information. A typical example is the relation between a data-table and a graph. One can also argue that the relation between a particular specific word and its entry in a dictionary belongs to this type. In the model worked out by Harmsze, the latter kind of relation is part of discourse relations (see below). This approach is warranted if such is a link is added to clarify a term with the help of a dictionary or an encyclopaedia. In a pure organisational framework, one can think of a series of equivalent representations of the same information. For example, acetylsalicylic acid (regular chemical name) = 2 actoxybenzoic acid (formal chemical name) = C9H804 (chemical formula) = Aspirin (trade name) = 2D picture of structure = 3D representation of structure.

1.6- Administrative relations that relate the various modules with the various kinds of meta-information collected in the module Meta-Information.

2- *Scientific discourse relations*

In [Har99] and [Har00], a distinction between two classes of discourse relations are given. One class is based on the communicative function. The author may want to argue or elucidate something in order to increase the reader's understanding or acceptance of the presented information. Hence, the author links the module to another module where the supportive information is given. In that link, the communicative function of the target with respect to its source can be made explicit. The different types of relations based on the communicative function are elucidation (which can be further expanded into clarification and explanation) and argumentation (which can be split into refutation or support of a standpoint). Important to note is that a clear distinction is made between links that aim to increase the reader's understanding and links that aim to increase to reader's acceptance of the presented material.

A second class exists of content relations that can be defined next to elucidatory and argumentative relations. In particular we defined, dependency, elaboration, synthesis, comparison and causality. The elaborate scheme thus obtained is the result of an inductive process and presents the relations that are the most relevant in the corpus. An important issue for further research is how the various relations based on the communicative function and the content relations can combine to form specific combinations that are useful in a particular domain.

A different approach might be based on Garssen [Gar97, Gar98], who showed that almost all so-called argumentation schemes discussed in the literature, that are schemes in which the acceptability of the premise that is explicit in the argumentation is transferred to the standpoint, can be divided into three categories that are clearly demarcated, homogeneous, mutually exclusive, and non-superfluous. These three relational categories, can be used in speech communication in various ways: as a description, as an argument and as a clarification or explanation. In the case of argumentation, a dispute has to be resolved; in the case of a clarification, we need to enhance the understanding of a statement.

2.1- Causal relations are relations where there is a causal connection between premise and conclusion (or between explanans and explanandum). This kind of relation exists between a statement or a formula and an elaborate mathematical derivation. Obviously, the usage of the causal relation as an argument and as an explanation, lie close together.

2.2- Comparison relations are relations where the relation is one of resemblance, contradiction or similarity. The analogue is a typical subtype. Comparisons used as argument are well-known phenomena, such as with the comparison of measured data from, e.g., the module Treated Results with theoretical predictions that fit within certain acceptable boundaries. We can also think of similarity relations, where results of others on similar systems are compared to emphasise agreement or disagreement. In the use of an elucidation, we can think of the relation between the description of a phenomenon and a known mechanical toy model. Furthermore, the link between a text and an image that illustrates the reasoning or results belong to this category. Another example is the suggestion that a drug that is effective in curing a particular ailment, might also help against look-alike symptoms, or the suggestion that look-alike physiological phenomena might have the same illness as the source.

2.3- Symptomatic relations, which are of a more complicated nature. Here we deal with relations where a concomitance exists between the two poles. This category is more heterogeneous than the other two. This kind of relation can be based on a definition or a value judgement such as the role of a specific feature that serves as a sufficiently discriminatory value to warrant a conclusion. We can think of a relation between the textually described results and a picture in which a specific feature, like a discontinuity in a graph, is used to declare a particular physical effect present or not.

Obviously next to these three classes, we still need the notion of relations that represent dependency of information in a line of reasoning and elaboration to cater for various levels of readership (e.g., extra information for the freshman reader).

5) Discussion

In the above, it has been argued that in an electronic environment, scientific papers can cease to be most basic independent self-contained entities. Instead of that, we suggest that the scientific communication of the future could be characterised by a network of modules defined by their conceptual function and other classes of identification.
This way, information can be shared by various authors, and in reporting new science, only those modules have to be written that present new information or insights. Next to that, commentaries to existing trails of modules can be added as independent entities. Such commentaries can be of different kinds: as new data in the form of a new module Raw Data, a new Theoretical Method or an independent Interpretation of other peoples work. Essential for such a system is the existence of a limited but sufficient number of non-overlapping categories of modules, and equally important, categories of relations that are expressed in labelled hyper-links. As with the design of all knowledge-related systems, the conceptualisation of the parameter space is very difficult.

The attempt made in this work is to try to keep the main descriptor classes (or coordinates) as domain-independent as possible, with the clear understanding that domain-specific identification in the form of keywords is one of the system's coordinates. In the field of knowledge-based systems, the term Ontology is now a fashionable term to describe the kind of systems based on both the semantics of every notion and the rules of using the primary terms in modelling a domain. Though the term is used in different ways [Vic97], the important notion is that contrary to IR, we deal with rules and relations as well. Many systems are built with the aim of assisting in the extraction of information. In our approach, we try to design a new input system, that, if successful, will substantially enhance the subsequent storage, retrieval and annotation capability of electronic publishing. In that process, we try to restrict the granularity of self-contained units of a well-defined conceptual meaning for which the development of a new writing style is important. Not the length, but the coherence of the message is the essence. This aim of conceptual granularity induces the need for a taxonomy of relations that is based on the real use of linguistic communication next to pure logical and structural approaches. Systematic classification of speech communication, and argumentation in particular, is therefore a valuable tool. A plethora of rhetorical indicators cannot define the necessary discriminatory coordinates, just as a thesaurus of keywords is insufficient to learn about the real content of a work. Our approach is bottom-up, given a conceptual level of granularity. In that sense, it has elements in common to the Plinius project in material science where also a bottom-up ontology is proposed [Vet98].

6.1) Findings

Our aim to develop a new framework for scientific communication based on the intrinsic characteristics of distributed electronic storage, has led us to the design of a modular framework. In this framework, a new granularity of scientific information is suggested, based on the conceptual content of the message. As every level of aggregation of information demands proper structuring, we introduced classes of relations that express themselves in hyperlinks connecting the modules or parts thereof. We provide a general systematic and coherent model that can be tailored to specific domain requirements. In the *Communication in Physics Project*, we showed that an application in experimental molecular physics is easily feasible. In this paper, we outlined the background and concepts in order to bring the discussion onto a more interdisciplinary level.

6.2) Arrows to further research

It goes without saying that much more in-depth analyses as well as user-surveys are necessary to feel safe with a minimum set of module and link descriptors. Further work is especially needed in the stratification of the discourse relations, and in particular, the possible connections between the communicative and content relations. The important issue is, that we need structural relation taxonomies and not arbitrary lists of rhetorical semantic indicators. Best practice would be to build a domain-specific experimental writing environment with an emphasis on authoring tools for reasonably large scale tests, where authors cast their publications in modular form. In such an authoring system, authors must be able to link the different modules or parts thereof with the help of a preset collection of relations. Analyses of such a practice, the problems and the results for the proper understanding by the ultimate readers must then provide the model with concrete feedback and suggestions for expanding the available types of relations.

This type of a more systematic and comparative study are needed in different domains. A good suggestion could be large conference proceedings in an experimental field, or a comprehensive work with lots of data.

1.6) Acknowledgements

Helpful discussions with Maarten van der Tol, Francisca Snoeck Henkemans, Anita de Waard and Keith Jones are gratefully acknowledged. The *Communication in Physics Project* is financially supported by the Foundation Physica, the Shell Research and Technology Centre Amsterdam, the Royal Dutch Academy of Sciences, the Royal Library and Elsevier Science BV.

1.7) References

[Bar96] Lisa Baron, Jean Tague-Sutcliffe, and Mark T. Kinnucan. Labelled, typed links as cues when reading hypertext documents. JASIS 47 (12), 1996, 896-908.

[Der89] Steven J. DeRose. Expanding the notion of links. Hypertext '89 Proceedings, November 1989, 249-257.

[Eam94] William Eamon. Science and the secrets of nature: Books of secrets in medieval and early modern culture. Princeton. Princeton Univ. Press, 1994.

[Eis79] Elizabeth. L. Eisenstein. The printing press as an agent of change: communications and cultural transformations in early modern Europe. 2 Volumes. Cambridge: Cambridge Univ. Press, 1979.

[Eis83] Elizabeth. L. Eisenstein. The printing revolution in Early modern Europe. Cambridge: Canto Cambridge Univ. Press, 1996.

[Feb76] Lucien Febvre and Henri-Jean Martin. The coming of the book. The impact of printing 1450-1800. London: Verso Press, 1976.

[Gar97] Bart Garssen. Argumentatieschema's in pragma-dialectisch perspectief. Een theoretisch en empirisch onderzoek. PhD thesis University of Amsterdam. Amsterdam: IFOTT Vol.32, 1997.

[Gar98] Bart Garssen. The nature of symptomatic argumentation. In: Frans H. van Eemeren, Rob Grootendorst, J Anthony Blair, Charles A. Wilards (eds.). Proceedings of the 4th International Conference of the International Society for the Study of Argumentation, Amsterdam, June 16-19 1998. Amsterdam: SICSAT, 1999.

[Har96] Frédérique Harmsze, Maarten van der Tol, Joost Kircz. Naar een modulair model voor natuurwetenschappelijke informatie in elektronische artikelen. In: K. van der Meer (red.). Informatiewetenschap 1996. Wetenschappelijke bijdragen aan de Vierde Interdisciplinaire Conferentie Informatiewetenschap. Delft, 13 december 1999.

[Har99] F.A.P. Harmsze, M.C. van der Tol and J.G. Kircz. A modular structure for electronic scientific articles. Contribution to the Conferentie Informatie Wetenschap 1999. CWI Amsterdam, 12 November 1999. On-line proceedings: http://www.cwi.nl/~lynda/WGI/info-wet1999/proceedings/. To be published.

[Har00] Frédérique Harmsze. A modular structure for scientific articles in an electronic environment. PhD Thesis, University of Amsterdam, Amsterdam, 2000. An electronic version of this thesis can be found at: http://www.wins.uva.nl/projects/commphys/papers/thesisfh/front.html

[Hav86] Eric. A. Havelock. The muse learns to write. Reflections on orality and literacy from antiquity to the present. New Haven: Yale Univ. Press, 1986.

[Kir91] Joost G. Kircz. Rhetorical structure of scientific articles: the case for argumentational analysis in information retrieval. Journal of Documentation, 47(4), 1991, 354-372.

[Kir98] Joost G. Kircz. Modularity: the next form of scientific information presentation? Journal of Documentation, 54(2), 1998, 210-235.

[Man88] W.C. Mann and S.A. Thompson, Rhetorical structure theory: towards a functional theory of text organization. Text, 8, 243-281.

[Man89] W.C. Man, C.M.I.M. Mattheissen and S.A. Thompson. Rhetorical structure theory and text analysis. Information Science Institute Research Report, Philadelphia: ISI/ RR-89-242, November 1989.

[Mcl62] M. McLuhan. The Gutenberg galaxy, the making of typographic man. Toronto. Univ. of Toronto Press, 1962.

[Mea98] A.J. Meadows. Communicating Research. San Diego: Academic Press, 1998.

[Ong82] W.J.Ong. Orality and literacy: the technologizing of the world. London: Methuen, 1982.

[RDF99] www.w3.org/TR/NOTE-rdf-simple-intro (last access 4-1-00).

[Tol99] M.C. van der Tol. The abstract as an orientation tool in modular electronic articles. In: Alfons Maes, Hans Hoeken, Leo Noordman, Wilbert Spooren (eds.) Document Design, Linking writer's goals to reader's needs. Proceedings of the First International Conference on Document Design. Tilburg University, 17-18 December, 1998. Tilburg, 1999, 175-186. An electronic version is at: http://www.wins.uva.nl/projects/papers/docdes/docdes.html

[Tri83] Randall Trigg. A network-based approach to text handling for the online scientific community. PhD dissertation Univ. of Maryland Tech. report, TR-1346, 1983. The taxonomy is given in chapter four which is available on: www.parc.xerox.com/spl/members/trigg/thesis/thesis-chap4.html.

[Rut00] Lloyd Rutledge, Brian Bailey, Jacco van Ossenbruggen, Lynda Hardman and Joost Geurts. Generating presentation constraints from rhetorical structure. Contribution to the Conferentie Informatie Wetenschap 1999. CWI Amsterdam, 12 November 1999. To be published. Also: http://www.cwi.nl/~lynda/WGI/info-wet1999/proceedings/ .

[Tuf83] Edward R. Tufte. The Visual Display of Quantitative Information. Cheshire: Graphics Press, 1983.

[Tuf90] Edward R. Tufte. Envisioning Information. Cheshire: Graphics Press, 1990.

[Vet98] Paul E. van der Vet and Nicolaas J.I. Mars. Bottom-up construction of ontologies. IEEE Transactions on Knowledge and Data Engineering 10, 1998, 513-526.

[Vic97] B.C. Vickery. Ontologies. Jnl. of Information Science, 23(4), 1997, 277-286.

[XML99] www.w3.org/TR/NOTE-xlink-req (last access 4-1-00).

MESH: an Object-Oriented Approach to Hypermedia Modeling and Navigation

Wilfried Lemahieu

Author information:

Dr. Wilfried Lemahieu

Department of Applied Economic Sciences
Katholieke Universiteit Leuven
Naamsestraat 69 B-3000 Leuven
Belgium
tel: + 32 16 32 68 86
fax: + 32 16 32 67 32
e-mail: wilfried.lemahieu@econ.kuleuven.ac.be

MESH: an Object-Oriented Approach to Hypermedia Modeling and Navigation

Wilfried Lemahieu

Abstract

This paper introduces the *MESH* approach to hypermedia modeling and navigation, which aims at relieving the typical drawbacks of poor maintainability and user disorientation. The framework builds upon two fundamental concepts. The *data model* combines established entity-relationship and object-oriented abstractions with proprietary concepts into a *formal hypermedia data model*. Uniform layout and link typing specifications can be attributed and inherited in a *static* node typing hierarchy, whereas both nodes and links can be submitted *dynamically* to multiple complementary classifications. In the *context-based navigation paradigm*, conventional navigation along static links is complemented by run-time generated *guided tours*, which are derived dynamically from the context of a user's information requirements. The result is a two-dimensional navigation paradigm, which reconciles complete navigational freedom and flexibility with a measure of linear guidance. These specifications are captured in a high-level, platform independent *implementation framework*.

1 Introduction

The hypermedia paradigm looks upon data as a network of *nodes*, interconnected by *links*. Whereas each node symbolizes a *concept*, a link not only stands for a *relation* between two items, but also explicitly assumes the semantics of a navigation path, hence the quintessential property of *navigational data access*. Their inherent flexibility and freedom of navigation raises hypermedia systems as utterly suitable to support user-driven exploration and learning. Therefore, hypermedia data retrieval embraces a notion of *location*. Data accessibility depends on a user's position in the network, denoted as the *current node* [Lucarella, 1990]. Manipulation of this position gradually reveals links to related information.

Unfortunately, due to inadequacy of the underlying data models, most hypermedia technologies suffer from severely limited maintainability. Moreover, the explorative, non-linear nature of hypermedia navigation imposes a heavy processing load upon the end user, referred to as *cognitive overhead* [Conklin, 1987]. The stringent problem of cognitive overhead effecting into user disorientation and losing one's chain of thought is known as the 'lost in hyperspace' phenomenon [Nielsen, 1990]; [Hammond, 1993]. Disorientation is further increased by the sense of *fragmentation* that is induced by scattering information over numerous separate nodes [Thüring et al., 1995].

This paper overviews the *MESH* hypermedia framework as deployed in [Lemahieu, 1999a], which proposes a structured approach to both data modeling and navigation, so as to overcome said maintainability and user disorientation problems. *MESH* is an acronym for *Maintainable, End user friendly, Structured Hypermedia*. The text, an extended version of [Lemahieu, 1999b], is partitioned according to *MESH*'s fundamental concepts. To start with, the *object-oriented hypermedia data model* is portrayed. The next section is dedicated to the *context-sensitive navigation paradigm*. A subsequent section translates these blueprints into a high-level *implementation framework*, specified in an abstract and platform independent manner. A last section makes comparisons to related work and formulates conclusions.

2 *An object-oriented hypermedia data model*

2.1 Introduction

The benefits of data modeling abstractions to both orientation and maintainability were already acknowledged in [Halasz, 1988]. They yield richer domain knowledge specifications and more expressive querying. Typed nodes and links offer increased consistency in both node layout and link structure [Thüring et al., 1991]; [Knopik & Bapat, 1994]. Higher-order information units and perceivable equivalencies (both on a conceptual and a layout level) greatly improve orientation [Thüring et al., 1995]; [Ginige et al., 1995]. Semantic constraints and consistency can be enforced [Garzotto et al., 1995]; [Ashman et al., 1997], tool-based development is facilitated and reuse is encouraged [Nanard & Nanard, 1995].

The first conceptual hypermedia modeling approaches such as *HDM* [Garzotto et al., 1993] and *RMM* [Isakowitz et al., 1995]; [Isakowitz et al., 1998] were based on the entity-relationship paradigm. Object-oriented techniques were mainly applied in *hypermedia engines*, to model functional behavior of an application's *components*, e.g. *Microcosm* [Davis et al., 1992]; [Beitner et al., 1995], *Hyperform* [Wiil & Leggett, 1997] and *Hyperstorm* [Bapat et al., 1996]. Along with the *Tower model* [De Bra et al., 1992], *EORM* [Lange, 1994] and *OOHDM* [Schwabe et al., 1996]; [Schwabe & Rossi, 1998a]; [Schwabe & Rossi, 1998b], *MESH* is the first approach where modeling of the *application domain* is fully accomplished through the object-oriented paradigm.

MESH's data model is based on concepts and experiences in the related field of database modeling, taking into account the particularities inherent to the hypermedia approach to data storage and retrieval. Established object-oriented modeling abstractions [Rumbaugh et al., 1991]; [Jacobson et al., 1992]; [Meyer, 1997]; [Snoeck et al., 1999] are coupled to proprietary concepts to provide for a *formal hypermedia data model*. While uniform layout and link typing specifications are attributed and inherited in a *static* node typing hierarchy, both nodes and links can be submitted *dynamically* to multiple complementary classifications. The data model provides for a firm hyperbase structure and an abundance of meta-information that facilitates implementation of an enhanced navigation paradigm.

2.2 The basic concepts: node types, layout templates and link types

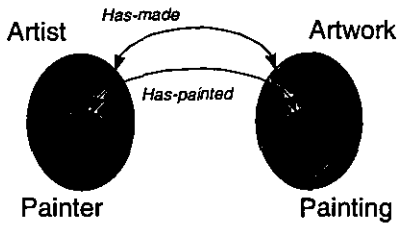
On a conceptual level, a *node* is considered a black box, which communicates with the outside world by means of its *links*. External references are always made to the node *as a whole*. True to the O.O. *information-hiding* concept, no direct calls can be made to its multimedia content. However, internally, a node may encode the intelligence to adapt its visualization to the *navigation context*, as discussed in section 0. Nodes are assorted in an inheritance hierarchy of *node types*. Each child node type should be compliant with its parent's definition, but may fine-tune inherited features and add new ones. These features comprise both node layout and node interrelations, abstracted in *layout templates* and *link types* respectively.

A *layout template* is associated with each level in the node typing hierarchy, every template being a refinement of its predecessor. Its exact specifications depend upon the implementation environment, e.g. as to the Web it may be *HTML* or *XML* based. Node typing as a basis for layout design allows for uniform behavior, onscreen appearance and link anchors for nodes representing similar real world objects.

A *link* represents a one-to-one association between two nodes, with both a semantic and a navigational connotation. A directed link offers an access path from its *source* to its *destination node*. Links representing similar semantic relationships are assembled into *types*. Link types are attributed to node types and can be inherited and refined throughout the hierarchy. Link type properties such as *domain*, *cardinalities* and *destination/inverse*[1] allow for enforcing constraints upon their instances.

[1] As discussed in detail in [Lemahieu, 1999a], a link type's *destination* is a derived property, defined as the *inverse link type's domain*.

These properties can be overridden to provide for stronger restrictions upon inheritance. E.g. whereas an **artist** node can be linked to any **artwork** through a ***has-made*** link type, an instance of the child node type **painter** can only be linked to a **painting**, by means of the more specific child link type ***has-painted***.



2.3 The use of aspects to overcome limitations of a rigid node typing structure

2.3.1 Definition of aspect descriptor and aspect type

The above model is based on a node typing strategy where node classification is total, disjoint and constant. The aspect construct allows for defining *additional* classification criteria, which are not necessarily subject to these restrictions. Apart from a single "most specific node type", they allow a node to take part in other secondary classifications that are allowed to change over time[1].

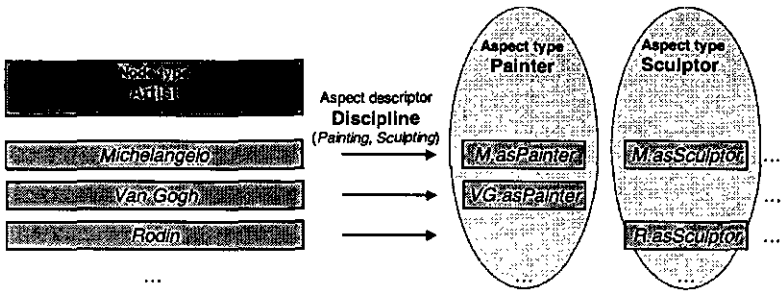
An *aspect descriptor* is defined as an attribute whose (discrete) values classify nodes of a given type into respective additional subclasses. In contrast to a node's "main" subtyping criterion, such aspect descriptor should not necessarily be *single-valued* nor *constant over time*. Aspect descriptor properties denote whether the classification is *optional/mandatory*, *overlapping/disjoint* and *temporary/ permanent.*

Each *aspect type* is associated with a single value of an aspect descriptor. An aspect type defines the properties that are attributed to the class of nodes that carry the corresponding aspect descriptor value. An aspect type's instances, *aspects*, implement these type-level specifications. Each aspect is inextricably associated with a single node, adding characteristics that describe a specific "aspect" of that node.

A node instance may carry multiple aspects and can be described by as many aspect descriptors as there are additional classifications for its node type. If multiple classifications exist, each aspect descriptor has as many values as there are subclasses to the corresponding specialization. Its cardinalities determine whether the classification is total and/or disjoint. As opposed to node types, aspects are allowed to be volatile. Hence, dynamic classification can be accomplished by manipulating aspect descriptor values, thus adding or removing aspects at run-time. Aspect types attribute the same properties as nodes: *link types* and *layout*. However, their instances differ from nodes in that they are not directly referable. An aspect represents the *same real-world object* as its associated node and can only be visualized as a subordinate of the latter.

E.g. to model an **artist** that can be skilled in multiple disciplines, a non-disjoint aspect descriptor *discipline* defines the **painter** and **sculptor** aspect types. Discipline-specific node properties are modeled in these aspect types, such that e.g. the **Michelangelo** node features the combined properties of its **Michelangelo.asPainter** and **Michelangelo.asSculptor** aspects.

1 We deliberately opted for a single inheritance structure, however, aspects can provide an elegant solution in many situations that would otherwise call for multiple inheritance, much like *interfaces* in the *Java* language. See [Lemahieu, 1999a] for further details.



2.3.2 Aspect types as node type building blocks

Node type properties (i.e. layout and link types) can be *delegated* to aspect descriptors, such that they can be inherited and overridden in each aspect type that is associated with one of the descriptor's values. An aspect type's *layout* template refines layout properties that are delegated to the corresponding aspect descriptor. Link types delegated to an aspect descriptor can be inherited and overridden as well. In addition, each aspect type can define its own supplementary link types. The inheritance/overriding mechanism is similar to the mechanism for supertypes/subtypes, but because an aspect descriptor can be multi-valued, particular care was taken so as to preclude any inconsistencies[1].

Aspect types themselves are node type properties that can be inherited and overridden across the node type hierarchy. The *aspect descriptor* is used as a vehicle for the inheritance of aspect types. This ability yields the opportunity to use aspects as real building blocks for nodes. Link types and layout definitions pertaining to a single "role" a node may have to play, can now be captured into one aspect type. If the corresponding aspect descriptor is attributed at a generic level in the node hierarchy, the aspect type can be inherited where necessary by more specific node types. This allows for the modeling of a similar 'aspect' in otherwise completely unrelated node types. Node types can be 'assembled' by inheriting the proper aspect types, complemented by their own particular features. Hereby, different aspects associated with the same node instance can have different editing privileges, such that updating multimedia *content* can be delegated to different parties.

2.4 Link typing and subtyping

2.4.1 Introduction

In common data modeling literature, subtyping is invariably applied to *objects*, never to *object interrelations*. If additional classification of a relationship type is called for, it is *instantiated* to become an object type, which can of course be the subject of specialization. However, as for a hypermedia environment, node types and link types are two separate components of the data model with very different purposes. It would not be useful to instantiate a link type into a node type, since such nodes would have *no content* to go along with them and thus each instance would become an 'empty' stop during navigation.

This section demonstrates how specialization semantics can be enforced not only upon node types, but also upon the *link types*. A sub link type will model a type whose set of instances constitutes a subset of its parent's, and which models a relation that is more specific than the one modeled the parent.

2.4.2 Definition and domain of a sub link type

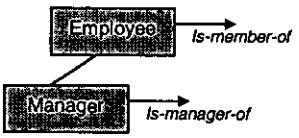
A link instance is defined as a source node - destination node tuple (n_s, n_d). Tuples for which this association represents a similar semantic meaning are grouped into link types.

[1] See [Lemahieu, 1999a] for further details.

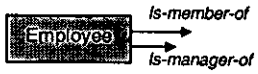
A link type defines instances that comply with the properties of the type and is constrained by its *domain*, its *cardinalities* and its *inverse link type*. The *domain* of the link type is the data type to which the link type is attributed. This can be either a node type or an aspect type.

If L_c is a sub link type resulting from a specialization over L_p, the set of (n_s, n_d) tuples defined by L_c is a subset of the one defined by L_p. Such specialization is called *vertical* if it is the consequence of a parallel classification over the link types' *domain*, denoting that the sub link type is attributed at a 'lower', more specific level in the node typing hierarchy than its parent. If L_c and L_p share the same domain, L_c can still define a subtype of L_p in the case where L_c models a more restricted, more specific kind of relationship than L_p, independently of any node specialization. Both parent and child link type are attributed at the same level in the node type hierarchy, hence the term *horizontal* specialization.

E.g.



***Vertical* link specialization**



***Horizontal* link specialization**

2.4.3 Overriding link type properties

Apart from the domain, a link type's cardinalities and inverse can be overridden as well upon specialization. The *cardinalities* determine the minimum and maximum number of link instances allowed for a given source node. *MESH* presents a formal overriding mechanism, wherein particular care is taken so as not to violate the parent's constraints, particularly in case of a non-disjoint classification. For further details we refer to [Lemahieu, 1999a].

The *inverse link type* is the *most specific* link type that encompasses all of the original link type's tuples, with reversed source and destination. There are two possibilities. If the 'inverse-of' relationship is mutual, we speak of a *particular* inverse, notation: $L \leftrightarrow Inv(L)$. If this is not the case, we speak of a *general* inverse, notation: $L \rightarrow Inv(L)$. A *particular inverse* models a situation where two link types are *each other's* inverse. Not counting source and destination's sequence, the two link types represent the same set of tuples, e.g. **employee.*is-member-of*** $\leftrightarrow$ **department.*members***. The term *particular inverse* is used because no two link types can share the same particular inverse.

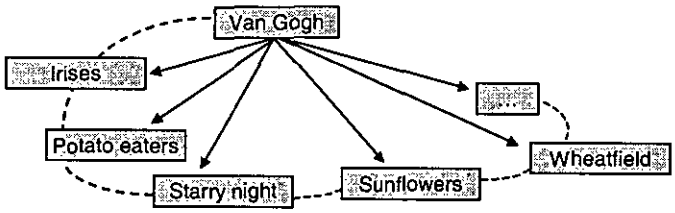
A child link type can override its parent's inverse with its own particular inverse, which is to be a subtype of the parent's inverse: **employee.*is-manager-of*** $\leftrightarrow$ **department.*manager***. However, if no suitable particular inverse exists for a given child link type, it has to inherit its parent's inverse as a *general inverse*, without overriding. Hence a *general* inverse can be shared by multiple link types with a common ancestor, e.g. **employee.*is-manager-of*** $\rightarrow$ **department.*members*** and **employee.*is-clerk-of*** $\rightarrow$ **department.*members***.

Link types are deemed extremely important, as they not only enforce semantic constraints but also *interface* between nodes, such that these can be coded and updated independently of one another. Moreover, they provide the basis for *context-sensitive node visualization*, as discussed in section 0.

3 *The context-based navigation paradigm*

3.1 Linearity and guided tours

To highlight the advantages of hypermedia navigation, comparisons are often made to books. Books are said to be *linear* information systems; their pages are organized uni-dimensionally, in a fixed order. Hypertext offers the possibility to break through this linear constraint and organize data in more complex structures, to be accessed following different possible paths, depending on the user's preferences and interests. Cognitive overhead, however, is significantly lower in a linear structure, be it at the cost of navigational freedom. Linearity provides a leading thread that facilitates orientation and prevents the reader from getting lost [Jonassen, 1990]. The latter is acknowledged in [Trigg, 1988]; [Nielsen, 1990], where linearity is re-introduced in so-called *guided tours*, chaining together all nodes pertaining to a common subject with *forward/backward* links. E.g. the typical hypermedia links (represented as arrows) between **Van Gogh** and each of his **paintings** can be complemented by a *guided tour* (represented as dotted lines) along these **paintings**.



Unfortunately, such hard-coded guided tours have proven to be inflexible and difficult to maintain. Moreover, they introduce a measure of redundancy into the hyperbase, as a guided tour typically reflects a communal property among its participating nodes. However, the property of '*being painted by the same artist*' is already established within the respective links from each **painting** to its **artist**. Thus, it would be possible to infer this knowledge and generate such guided tour *at run-time*, without burdening hyperbase maintainability.

MESH builds upon its data model to reconcile navigational freedom with the ease of linear navigation. Its intended navigation mechanism is that of an "intelligent book", which is to provide a disoriented end user with a sequential path as a guidance. Such guided tour is not static, but is adapted dynamically to the *navigation context*. In addition, a node is able to tune its *visualization* to the context in which it is accessed, hence providing the user with the most relevant subset of its embedded multimedia objects.

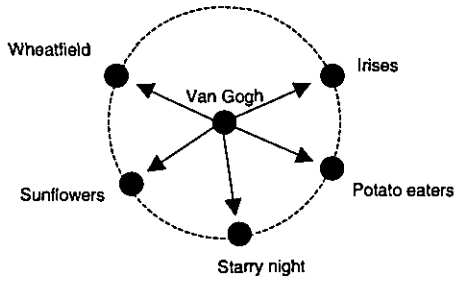
3.2 A guided tour as derived from the current context

In conventional hypermedia applications, the *current node* is the only variable that determines which information is accessible at a given moment; navigation is only possible to nodes that are linked to this current node. Its value changes with each navigation step as it represents the immediate focus of the user's attention. *MESH* introduces the *current context* as a second, longer-term variable that 'glues' the various visited nodes together and provides a background about which common theme is being explored. The current context is defined as the combination of a *context node* and a *context link type*. The context node represents the subject around which the user's broader information requirements 'circle'. The nature of the relationship involved is depicted by the context link type.

A guided tour derives from the current context. Therefore, *MESH* discriminates between *direct* and *indirect* links. A direct link represents a lasting relation between two nodes. Direct links are typed and reflect the underlying conceptual data model. Because they are permanent and context-independent, they are stored explicitly into the hyperbase and are always valid. E.g. the node **Sunflowers** is directly linked to the **Van Gogh** node.

An *indirect* link between two nodes indicates that they share relevancy to a common third node. The latter denotes the *context* within which the indirect link is valid. As indirect links not only reflect the data model, but also depend on a run-time variable, the *current context*, they cannot be stored within the hyperbase. They are to be created *dynamically* at run-time, as inferred from a particular context. E.g. an indirect link between **Sunflowers** and **Wheatfield** is only relevant when exploring information related to **Van Gogh**.

A *guided tour* is defined as a path of *indirect* links along all nodes relevant to the current context. These nodes are directly linked to the context node (through instances of the context link type) and indirectly to their predecessor and successor in the tour. As they are chained into a linear structure, a logical order should be devised in which the subsequent tour nodes can be presented to the user. The most obvious criterion is in alphabetical order of a *node descriptor* field. More powerful alternatives are discussed in [Lemahieu, 1999a]. E.g. the context **Van Gogh.*has-painted*** yields a guided tour among the nodes {**Irises, Potato eaters, Starry night, Sunflowers, Wheatfield,** ...} with **Van Gogh** as the *context node* and ***has-painted*** as the *context link type*.

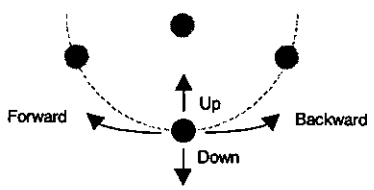


Current context: **Van Gogh.*has-painted***

Note that the discrepancy between *guided tour* and *context* can be compared to the traditional duality in representing a circle either through the points on its *circumference*, or through its *center* and a *radius*. Guided tours are not stored within the hyperbase as *an enumeration of participating nodes*, but are calculated at run-time from the *current context*. Although sequential by nature, such tours do not restrict the user's navigational freedom, as long as sufficient flexibility is offered in choosing which tour to follow. The linearity lies in 'following' the tour. The freedom lies in starting one.

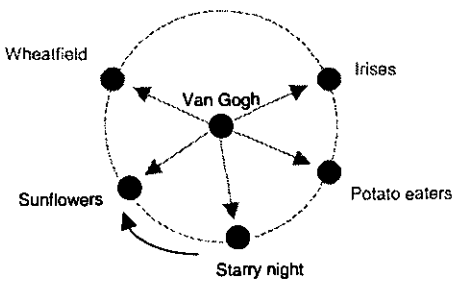
3.3 Navigational actions

Navigational actions can be classified according to two dimensions. First, there is moving *forward* and *backward* within the current tour, along indirect links. Second, and orthogonal to this, there is the option of moving *up* or *down* along direct links, closer to or further away from the session's starting point. Additionally, one can distinguish between actions that change the current context and actions that only influence the current node.



3.3.1 Moving forward/backward within the current tour

Moving forward or backward in a guided tour along indirect links, results in the node following/preceding the current node being accessed to become the new current node. The current context is unaffected.



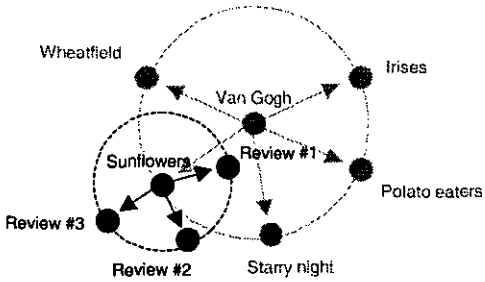
Current context: **Van Gogh.*has-painted***
Current node: **Starry night ⇒ Sunflowers**

3.3.2 Moving up/down

Moving down implies an action of 'digging deeper' into the subject matter, moving away from the starting point. This is accomplished through selection from the current node of either a direct link type or link instance. In the case of a *unique* destination node, the result is the latter node being accessed. In the case of a *set* of destination nodes, the outcome is a new 'nested' tour being started.
In complete analogy to traditional hypermedia navigation, selection of a link instance l from a given source node n_s results in its unique destination node n_d being accessed: $n_s.l := \{n_d \mid l = (n_s, n_d)\}$. E.g. selection of the link (**Sunflowers, National Gallery**) from the current node **Sunflowers**, induces an access to the node **National Gallery**; **Sunflowers**.(Sunflowers, National Gallery) := {**National Gallery**}.

However, *MESH* aggregating single *link instances* into *link types*, yields the opportunity of anchoring and consequently selecting a *complete link type* from a given source node. Selection of a link type L from a source node n_s yields a set of all destination nodes n_d of tuples representing link instances of L with n_s as the source node, i.e. all nodes that are linked to the current node by the selected link type: $n_s.L := \{n_d \mid (n_s, n_d) \in L\}$. Depending on maximum cardinality of the link type, the resulting set may contain multiple destination nodes. E.g. with **Sunflowers** as the current node, selection of the link type ***reviews*** generates a *collection* of nodes to-be-accessed: **Sunflowers.*reviews*** := {**review#1**, **review#2**, **review#3**, ...}.

The result of such action is a *context change*: a new context emanates, resulting in new indirect links. A new tour is generated, nested within the former, according to this new context. The current node **Sunflowers** is denoted as the new context node. The non-unique link type ***reviews*** defines the context link type, which yields a new *nested* tour: **Sunflowers.*reviews***. The first review is accessed to become the new current node. Such *context change* reflects the user's decision to concentrate on the current node as a new topic of interest. All indirect links are destroyed and redefined around this new context.



Current context: **Van Gogh.*has-painted* ⇒ Sunflowers.*reviews***
Current node: **Sunflowers ⇒ Review#1**

Hence contexts, and consequently guided tours, can exist in layers. As such, it is possible to 'delve' into a subject and have multiple *open* tours, nested within one another, where the context node of one tour is the current node of the tour it is nested in. Navigation along indirect links is invariably carried out within the "deepest", i.e. most recently started tour. Continuing a tour on a higher level is only possible if all tours on a lower level have been either completed or disbanded.

The latter is accomplished by *moving up*, which reverses the latest *move down* action. If the latter involved a context change, the move up action results in the reestablishment of the previous context and the cancellation of the tour generated through this most recent link type selection. The previous context's context node and indirect links are restored. The most recent context node (**Sunflowers** in the example) again becomes the current node.

The practice of node and link typing allows for casting navigational actions to a whole *class* of nodes, regardless of the actual instance they are applied to. Hereby, selections of link *types* that exist at a sufficiently high level of abstraction can be imposed upon every single node belonging to a tour. E.g. in the context of **Van Gogh.*has-painted***, a **painting**#x.***reviews*** selection can be issued once on *tour* level, with additional (nested) tours being generated automatically for each node participating in the **Van Gogh.*has-painted*** tour. If these tours in their turn include navigational actions on type level, a complex navigation pattern results, which can be several levels deep. Again, *forward* and *backward* links always apply to the current tour, i.e. to the open tour at the 'deepest' level. In addition, the abstract navigational actions and tour definitions sustain the generation of very compact tree-shaped overviews and maps of complete navigation sessions[1]. In this respect, the *move up* and *move down* actions indeed correspond to moving up or down in the graph. The represented information can also be *bookmarked*, i.e. bookmarks not just refer to a single node but to a complete navigational situation, which can be resumed at a later time.

4 A generic application framework

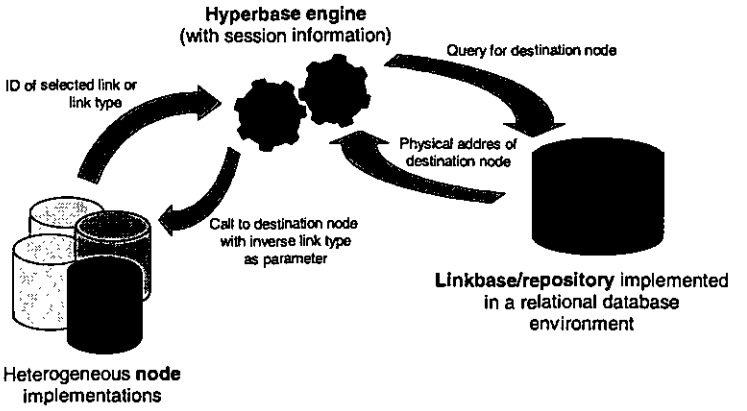
The *information content* and *navigation structure* of the nodes are separated and stored independently. The resulting system consists of three types of components: the *nodes*, the *linkbase/repository* and the *hyperbase engine*. In [Lemahieu, 1999a], a platform-independent implementation framework was provided, but all subsequent prototyping is explicitly targeted at a *Web* environment.

A node can be defined as a static page or a dynamic object, using e.g. *HTML* or *XML*. Its internal content is shielded from the outside world by the indirection of link types playing the role of a node's *interface*. Optionally, it can be endowed with the intelligence to tune its reaction to the *context* in which it is accessed by integrating the node type's set of attributed link types as a parameter in its layout template's presentation routines. Upon activation, a node is provided with the link type by which it was accessed. Consequently, the multimedia objects that are most relevant to this particular link type can be made current, hence the so-called *context-sensitive visualization* principle. This allows for different views to be defined over the same information, much like the *city* construct in the *Tower* model.

Since a node is not specified as a necessarily searchable object, linkage information cannot be embedded in a node's body. Links, as well as meta data about node types, link types, aspect descriptors and aspects are captured within a searchable *linkbase/repository* to provide the necessary information pertaining to the underlying hypermedia model, both at design time and at run-time. This repository is implemented in a relational database environment. Only here, references to physical node addresses are stored, these are never to be embedded in a node's body. All external references are to be made through location independent *node ID*'s.

The *hyperbase engine* is conceived as a server-side script that accepts link (type) selections from the current node, retrieves the correct destination node, keeps track of session information and provides facilities for generating maps and overviews. Since all relevant linkage and meta information is stored in the relational DBMS, the hyperbase engine can access this information by means of simple, pre-defined and parameterized *database queries*, i.e. without the need for searching through *node content*.

[1] See [Lemahieu, 1999a] for further details.



5 *Conclusions*

5.1 Data modeling and authoring

Ohter hypermedia approaches such as *EORM*, *RMM*, *HDM* and *OOHDM* are also based on conceptual modeling abstractions, either through E.R. or O.O techniques. Among these, *OOHDM* is the only methodology to incorporate a subtyping and inheritance/overriding mechanism. However, subtyping modalities are not explicitly stipulated. Rather, they are borrowed from *OMT* [Rumbaugh et al., 1991], a general-purpose object-oriented design methodology.

MESH deploys a proprietary approach, specifically tailored to hypermedia modeling, where *structure* and *relationships* prevail over *behavior* as important modeling factors. Its full O.O. based data modeling paradigm should allow for hypermedia *maintenance* capabilities equaling their database counterpart; with unique object identifiers, monitoring of integrity, consistency and completeness checking, efficient querying and a clean separation between authoring *content* and *physical hyperbase maintenance*. *MESH* is the only approach to formulate specific rules for inheriting and overriding layout and link type properties, taking into account the added complexity of plural (possibly overlapping and/or temporal) node classifications. Links are treated as first-class objects, with link types being able to be subject to multiple specializations themselves, not necessarily in parallel with node subtyping. Authoring is greatly facilitated by O.O. features such as inheritance and overriding, class properties and layout templates that allow for high-level specification and lower-level refinement of node properties. Links can be anchored on *type* level, independently of actual node and link instances. Semantics attributed within the data model permit the automated type checking and integrity constraints we have grown accustomed to in a database environment. Dangling links and inconsistent link attributions can already be detected during the design phase. Node design is further enhanced by links and layout properties being automatically suggested. For that purpose, a design tool can retrieve the necessary information from the meta knowledge stored within the hyperbase. Finally, it is clear that a model-based approach in general facilitates information sharing, reuse, development in parallel, etc.

5.2 Navigation and orientation

Apart from the obvious benefit of a well-maintained hyperbase, *typed links* should permit a better comprehension of the semantic relations between information objects. The use of *higher-order* information units and the representation of collections of nodes as (source node, link type) combinations, induces a stronger sense of structure. A *node typing hierarchy* with consistent layout and user interface features, reflecting similarities between nodes, is to reduce both cognitive overhead and the impression of fragmentation. The *context* concept, as a representation of what various nodes have in common, will diminish fragmentation, but is to remedy cognitive overhead as well, as the linear guided tours improve a user's sense of position and his ability to ascertain his navigational options. Through the specification of navigational actions *on tour level*, complex navigation patterns can be applied to all nodes in a tour without additional effort.

A further decrease in fragmentation and cognitive overhead is obtained by making node *visualization* dependent upon the context in which a node is accessed. The abundance of meta-information as node, aspect and link types allows for enriching *maps* and *overviews* with concepts of varying granularity. A final benefit is the ability to *bookmark a complete navigational situation* in an utterly compact manner, with the possibility of it being resumed later on, from the exact point where it was left.

Set-based hypermedia paradigms such as *CHM* [Duval et al., 1995a]; [Duval et al., 1995b], the *HM-Data Model* [Maurer & Scherbakov, 1992]; [Srinivasan, 1995] and *Hyper-G* [Andrews et al., 1995a]; [Andrews et al., 1995b] equally provide inherent support for navigation in two orthogonal planes; *inside a collection* and *across collection boundaries*. Hereby, their *current container* and *current member* concepts are comparable to *MESH*'s *current context* and *current node* respectively. A drawback, however, is that they *do* severely limit navigational freedom. Moreover, they lack a firm underlying data model with typed node interrelations. Likewise, the opportunity of defining abstract navigational actions on tour level is a feature that is exclusive to *MESH*.

EORM, *RMM*, *HDM* and *OOHDM* also feature specific topologies such as *guided tours*, *indexes* etc. A fundamental difference is that these are conceived as explicit *design components*, requiring author input for query definitions, node collections and forward/backward links. In *MESH*, neither guided tours nor indexes require any maintenance nor design effort, as the author is not even engaged in their realization. They are generated at run-time upon user request, not to restrict his freedom, but merely to facilitate navigation and support orientation.

5.3 Ongoing research issues

In the current stage of development, data modeling and navigation have been favored over internal node design. Indeed, each object from the conceptual domain model is to be translated to an *active* node, which is responsible for its own context-dependent visualization and interaction with the user. In this respect, node types can be seen as *tower objects* [De Bra et al., 1992], with a node in itself bearing a description on different levels, e.g. structural, semantic, presentational etc. For that purpose, ongoing research is explicitly targeted at a *Web* environment. As to the latter, the *XML* standard could be of utter importance.

A related issue is the support for *continuous media types*. Media type specific manipulations are currently considered to be internal node properties, beyond the scope of the model. However, synchronization of multiple audio and video tracks requires facilities for inter node and intra node timing constraints [Hardman et al., 1994]. These cannot be enforced in *MESH* at present. Moreover, the constraint of *only a single current node at a time* may be too restrictive, as suggested in [Hardman et al., 1993]; [De Bra et al., 1994]. Hereby, *MESH*'s *context node* concept can prove to be a valuable asset to allow for continuous media objects to keep playing "in the background", while other related nodes are visited in the same context.

References

[Andrews et al., 1995a] K. Andrews, F. Kappe and H. Maurer, The Hyper-G Network Information System, *Journal of Universal Computer Science, Vol. 1*, No. 4 (Apr. 1995)

[Andrews et al., 1995b] K. Andrews, F. Kappe and H. Maurer, Serving Information to the Web with the Hyper-G Network Information System, *Computer Networks and ISDN Systems, Vol. 27*, No. 6 (Apr. 1995)

[Ashman et al., 1997] H. Ashman, A. Garrido and H. Oinas-Kukkonen, Hand-made and Computed Links, Precomputed and Dynamic Links, *Proceedings of Hypertext - Information Retrieval - Multimedia (HIM '97)*, Dortmund (Sep. 1997)

[Bapat et al., 1996] A. Bapat, J. Wäsch, K. Aberer and J. Haake, An Extensible Object-Oriented Hypermedia Engine, *Proceedings of the seventh ACM Conference on Hypertext (Hypertext '96)*, Washington D.C. (Mar. 1996)

[Beitner et al., 1995] N. Beitner, C. Goble and W. Hall, Putting the Media into Hypermedia, *Proceedings of the SPIE Multimedia Computing and Networking 1995 Conference*, San Jose (Feb. 1995)

[Conklin, 1987] J. Conklin, Hypertext: An Introduction and Survey, *IEEE Computer Vol. 20*, No. 9 (Sep. 1987)

[Davis et al., 1992] H. Davis, W. Hall, I. Heath, G. Hill and R. Wilkins, MICROCOSM: An Open Hypermedia Environment for Information Integration, *Computer Science Technical Report CSTR 92-15* (1992)

[De Bra et al., 1992] P. De Bra, G. Houben and Y. Kornatzky, An Extensible Data Model for Hyperdocuments, *Proceedings of the fourth ACM European Conference on Hypermedia Technology (ECHT '92)*, Milan (Dec. 1992)

[De Bra et al., 1994] P. De Bra, G. Houben and Y. Kornatzky, A Formal Approach to Analysing the Browsing Semantics of Hypertext, *Proceedings of Computing Science in the Netherlands (CSN-94)*, Utrecht (1994)

[Duval et al., 1995a] E. Duval, H. Olivié, P. Hanlon and D. Jameson, HOME: An Environment for Hypermedia Objects, *Journal of Universal Computer Science Vol. 1*, No. 5 (May 1995)

[Duval et al., 1995b] E. Duval, H. Olivié and N. Scherbakov, Contained Hypermedia, *Journal of Universal Computer Science, Vol. 1*, No. 10 (Oct. 1995)

[Garzotto et al., 1993] F. Garzotto, P. Paolini, and D. Schwabe, HDM - A Model-Based Approach to Hypertext Application Design, *ACM Trans. Inf. Syst. Vol. 11*, No. 1 (Jan. 1993)

[Garzotto et al., 1995] F. Garzotto, L. Mainetti and P. Paolini, Hypermedia Design, Analysis, and Evaluation Issues, *Commun. ACM Vol. 38*, No. 8 (Aug. 1995)

[Ginige et al., 1995] A. Ginige, D. Lowe and J. Robertson, Hypermedia Authoring, *IEEE Multimedia Vol. 2*, No. 4 (Winter 1995)

[Halasz, 1988] F. Halasz, Reflections on NoteCards: Seven Issues for Next Generation Hypermedia Systems, *Commun. ACM Vol. 31*, No. 7 (Jul. 1988)

[Hammond, 1993] N. Hammond, Learning with Hypertext: Problems, principles and Prospects, *HYPERTEXT a psychological perspective*, C. McKnight, A. Dillon and J. Richardson Eds., *Ellis Horwood*, New York (1993)

[Hardman et al., 1993] L. Hardman, D. Bulterman and G. Van Rossum, Links in Hypermedia, the Requirements for Context, *Proceedings of the fifth ACM conference on Hypertext (Hypertext '93)*, Seattle (Nov. 1993)

[Hardman et al., 1994] L. Hardman, D. Bulterman and G. Van Rossum, The Amsterdam Hypermedia Model, *Commun. ACM Vol. 37*, No. 2 (Feb. 1994)

[Isakowitz et al., 1995] T. Isakowitz, E. Stohr and P. Balasubramanian, RMM, A methodology for structured hypermedia design, *Commun. ACM Vol. 38*, No. 8 (Aug. 1995)

[Isakowitz et al., 1998] T. Isakowitz, A. Kamis and M. Koufaris, The Extended RMM Methodology for Web Publishing, *Working Paper IS-98-18, Center for Research on Information Systems*, 1998 (Currently under review at ACM Trans. Inf. Syst.)

[Jacobson et al., 1992] I. Jacobson, M. Christerson, P. Jonsson and G. Övergaard, Object-Oriented Software Engineering, *Addison-Wesley*, New York (1992)

[Jonassen, 1990] D. Jonassen, Semantic net elicitation: tools for structuring hypertext, *Hypertext: State of the Art*, R. McAleese and C. Green Eds., *Intellect*, Oxford (1990)

[Knopik & Bapat, 1994] T. Knopik and A. Bapat, The Role of Node and Link Types in Open Hypermedia Systems, *Proceedings of the sixth ACM European Conference on Hypermedia Technology (ECHT '94)*, Edinburgh (Sep. 1994)

[Lange, 1994] D. Lange, An Object-Oriented design method for hypermedia information systems, *Proceedings of the twenty-seventh Hawaii International Conference on System Sciences (HICSS-27)*, Hawaii (Jan. 1994)

[Lemahieu, 1999a] W. Lemahieu, Improved Navigation and Maintenance through an Object-Oriented Approach to Hypermedia Modelling, *Doctoral dissertation (unpublished)*, Leuven (Jul. 1999)

[Lemahieu, 1999b] W. Lemahieu, MESH: an Object-Oriented Approach to Hypermedia Modeling and Navigation, *Proceedings of Informatiewetenschap '99*, Amsterdam (Nov. 1999)

[Lemahieu, 2000a] W. Lemahieu, A Context-Based Navigation Paradigm for Accessing Web Data, *Proceedings of the ACM Symposium on Applied Computing (SAC 2000)*, Como (Mar. 2000)

[Lemahieu, 2000b] W. Lemahieu, An Object-Oriented Approach to Conceptual Hypermedia Modeling, *Proceedings of IRMA 2000*, Anchorage (May 2000) (forthcoming)

[Lucarella, 1990] D. Lucarella, A Model For Hypertext-Based Information Retrieval, *Proceedings of the European Conference on Hypertext*, Versailles (Nov. 1990)

[Maurer & Scherbakov, 1992] H. Maurer and N. Scherbakov, The HM Data Model, *IIG Report*, Graz (1992)

[Meyer, 1997] B. Meyer, Object-Oriented Software Construction, Second Edition, *Prentice Hall Professional Technical Reference*, Santa Barbara (1997)

[Nanard & Nanard, 1995] J. Nanard and M. Nanard, Hypertext Design Environments and the Hypertext Design Process, *Commun. ACM Vol. 38*, No. 8 (Aug. 1995)

[Nielsen, 1990] J. Nielsen, The Art of Navigating Through Hypertext, *Commun. ACM Vol. 33*, No. 3 (Mar. 1990)

[Rumbaugh et al., 1991] J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy and W. Lorensen, Object Oriented Modelling and Design, *Prentice Hall*, Englewood Cliffs (1991)

[Schwabe et al., 1996] D. Schwabe, G. Rossi and S. Barbosa, Systematic Hypermedia Application Design with OOHDM, *Proceedings of the seventh ACM conference on hypertext (Hypertext '96)*, Washington DC (Mar. 1996)

[Schwabe & Rossi, 1998a] D. Schwabe and G. Rossi, Developing Hypermedia Applications using OOHDM, *Proceedings of the ninth ACM Conference on Hypertext (Hypertext '98)*, Pittsburgh (Jun. 1998)

[Schwabe & Rossi, 1998b] D. Schwabe and G. Rossi, An O.O. approach to web-based application design, *Draft* (1998)

[Snoeck et al., 1999] M. Snoeck, G. Dedene, M. Verhelst and A. Depuydt, Object-Oriented Enterprise modeling with MERODE, *Universitaire Pers Leuven*, Leuven (1999)

[Srinivasan, 1995] P. Srinivasan, Incorporating Intelligent Navigational Techniques to Hypermedia, *Proceedings of LAIR-MIPRO '95*, Opatija (May 1995)

[Thüring et al., 1991] M. Thüring, J. Haake, and J. Hannemann: What's ELIZA doing in the Chinese Room - Incoherent Hyperdocuments and how to Avoid them, *Proceedings of the third ACM Conference on Hypertext (Hypertext '91)*, San Antonio (Nov. 1991)

[Thüring et al., 1995] M. Thüring, J. Hannemann and J. Haake: Hypermedia and Cognition: Designing for comprehension, *Commun. ACM Vol. 38*, No. 8 (Aug. 1995)

[Trigg, 1988] R. Trigg, Guided Tours and Tabletops: Tools for Communicating in a Hypertext Environment, *ACM Trans. Office Inf. Syst. Vol. 6*, No. 4 (Oct. 1988)

[Wiil & Leggett, 1997] U. Wiil and J. Leggett, Hyperform: a hypermedia system development environment, *ACM Trans. Inf. Syst. Vol. 15*, No. 1 (Jan. 1997)

Op weg naar de virtuele informatieketen

Prof. John Mackenzie Owen (owen@hum.uva.nl)
Universiteit van Amsterdam

Inleiding

De wetenschappelijke informatievoorziening is in de afgelopen decennia als gevolg van de digitalisering drastisch veranderd. De technologische innovatie heeft geleid tot een infrastructuur die er heel anders uitziet dan aan het begin van de jaren tachtig van de twintigste eeuw. Die infrastructuur kan het beste worden beschreven met de omschrijving 'hybride'. Daarmee wordt aangegeven dat de instituties en functies binnen de wetenschappelijke informatieketen gebaseerd zijn op de nu eenmaal bestaande praktijk waarin *digitale* en *niet-digitale* informatie naast elkaar bestaan. Maar ondanks alle technologische veranderingen zijn veel kenmerken van de informatieketen toch hetzelfde gebleven. Een van die kenmerken is de manier waarop gespecialiseerde taken worden vervuld door afzonderlijke institutionele partijen, waaronder uitgevers en bibliotheken. Opvallend daarbij is dat de verantwoordelijkheid voor de wetenschappelijke informatievoorziening niet bij de wetenschappers berust, maar verdeeld is over de verschillende partijen in de informatieketen zonder dat er sprake is van een centrale regie. Volgens dit model is de universiteit verantwoordelijk voor de creatie van 'content', de uitgever voor de verpakking en distributie daarvan, en de bibliotheek voor ontsluiting, beheer en beschikbaarstelling. De onderscheiden taken en de instituties die ze uitvoeren, zijn ondanks alle technologische vernieuwing in hoge mate dezelfde gebleven.

In de Verenigde Staten kristalliseert zich momenteel onder de benaming 'digital library' een heel nieuw concept uit waarin sprake is van een nieuwe vorm van wetenschappelijke informatievoorziening. Deze nieuwe vorm is niet meer gebaseerd op de traditionele informatieketen, maar op een 'human-centered' cyclisch model van het informatiegedrag van wetenschappers.[1] Hierbij wordt de wetenschapper gezien als iemand die informatie en kennis produceert in de vorm van informatieobjecten.
Dit zijn objecten die steeds minder voldoen aan de kenmerken van traditionele documenten (zoals tijdschriftartikelen en monografieën): het gaat om tekstcorpora, dataverzamelingen, simulaties, modellen, beeldmateriaal en teksten, die gedistribueerd zijn op het netwerk, op verschillende manieren aan elkaar gekoppeld zijn en met elkaar kunnen interacteren. Om hiermee te kunnen werken, zijn systemen nodig voor distributie en opslag, inclusief lange-termijnarchivering. Ook heeft de gebruiker behoefte aan geavanceerde technieken voor navigatie en retrieval. Dit alles speelt zich af in een digitale netwerkomgeving: de 'digital library'.[2] Deze 'digital library' is niet een lokale organisatie van bibliothecarissen, maar een mondiale organisatie van wetenschappers. We zullen de 'digital library' in deze bijdrage daarom verder aanduiden als een 'virtuele informatieketen'.

Een dergelijk concept houdt waarschijnlijk ingrijpende innovaties, in dat en leidt tot de vraag hoe toekomstvast de huidige hybride informatieketen is waarin min of meer traditionele instituties de wereld van gedrukte publicaties met die van digitale informatie met elkaar hebben weten te combineren. Maar een ingrijpende verandering van de informatieketen houdt ook risico's in die kritische vragen oproepen. Leidt de virtuele informatieketen daadwerkelijk tot een toegevoegde waarde voor de wetenschapsbeoefening? Hoe komen we daar achter zonder ongewenste en wellicht onherstelbare effecten op te roepen? In welke mate moeten we in Nederland op zulke ontwikkelingen aansluiten en wellicht een voortrekkersrol vervullen? Wat kunnen die ontwikkelingen betekenen voor de bestaande instituties? In welke mate maken ze beleids- en gedragsveranderingen noodzakelijk in de wetenschappelijke wereld? Deze bijdrage pretendeert niet zulke vragen te beantwoorden, maar ze slechts aan de orde te stellen.

Technologische innovatie van de wetenschappelijke informatievoorziening

Een aanzienlijk deel van de technologische innovatie van de wetenschappelijke informatievoorziening is tot stand gekomen op basis van een aantal omvangrijke programma's voor onderzoek en ontwikkeling.[3] Voorbeelden van dergelijke programma's zijn:

- Nederland: *IWI-projecten*
- Europese Commissie: *Kaderprogramma's* III, IV, V
- ERCIM *Digital Library Initiative* (European Research Community for Informatics and Mathematics)
- UK: LIC/RIC *Digital Library Programme*
- UK: *Electronic Libraries Programme* (eLib)
- USA: *Digital Libraries Initiative* (DLI)
- USA: *National Digital Library Programme* (Library of Congress)
- Australia: projecten van de National Library of Australia
- Nieuw Zeeland: *Digital Library Project* van de University of Waikato
- Canada: *Canadian Initiative on Digital Libraries*

In deze programma's hebben met name de bibliotheken een belangrijke rol gespeeld als voortrekkers van de innovatie van de informatieketen als geheel. Daarom worden zulke programma's meestal aangeduid als digital (of electronic) libraries programme (of initiative).

Naast projecten uitgevoerd in het kader van de grote R&D-programma's zien we in toenemende mate ook R&D-activiteiten die op eigen initiatief en kosten door een of meer bibliotheken worden opgezet. De reden hiervoor ligt in het feit dat uitkomsten van de genoemde R&D-programma's - zoals de in Europa gebruikelijke 'hybride' bibliotheken (zie hierna) - zodanig uitontwikkeld zijn dat anderen zonder veel moeite van de resultaten van eerder onderzoek gebruik kunnen maken. Met andere woorden: het inrichten van (aspecten van) de huidige vorm van de digitale bibliotheek is geen R&D-inspanning meer, maar behoort onderhand tot de min of meer normale bedrijfsvoering.

In deze programma's hebben uitgevers in het algemeen geen dominante rol gespeeld, hoewel ze in beperkte mate wel aan projecten hebben bijgedragen, bijvoorbeeld door het leveren van digitale content. Ook de meer specifiek op de uitgeverswereld gerichte programma's van de Europese hebben op de wetenschappelijke uitgevers weinig aantrekkingskracht uitgeoefend. Het is dan ook niet verwonderlijk dat deze uitgevers relatief langzaam tot innovatie zijn gekomen, en in feite zijn blijven steken bij het in digitale vorm aanbieden van traditionele uitgeefproducten.

Resultaten van de R&D-programma's

Binnen de verschillende onderzoeksprogramma's zijn in het afgelopen decennium honderden projecten uitgevoerd die tot een scala van resultaten hebben geleid. Die resultaten liggen niet alleen op het gebied van (de toepassing van) nieuwe informatie- en communicatietechnologie. De spin-off van dergelijke programma's voor de deelnemende organisaties blijkt veel groter te zijn:

- Technische oplossingen: nieuwe systemen en producten voor gebruik binnen de informatievoorziening
- Nieuwe vormen van dienstverlening aan de gebruiker
- Standaards en protocollen voor de communicatie tussen systemen en de uitwisseling van informatie
- Nieuwe werkmethoden en 'best practices'
- Meer inzicht in het gedrag van gebruikers met betrekking tot de informatievoorziening in het algemeen en digitale informatie en diensten in het bijzonder
- Nieuwe werkrelaties in de informatieketen, bijvoorbeeld tussen bibliotheken onderling (ook internationaal) en tussen bibliotheken en uitgevers
- Aanzienlijke versterking van de kennis en vaardigheden op technisch, organisatorisch en communicatief gebied bij alle partijen in de informatieketen

Bij analyse van het onderzoek van de afgelopen jaren komen enkele algemene inzichten naar voren met betrekking tot de specifieke kenmerken van onderzoek en ontwikkeling op dit terrein. Daarbij is een viertal aspecten met name van belang gebleken:

- *Interoperabiliteit*: systemen moeten met elkaar kunnen samenwerken, en informatie moet kunnen worden verwerkt door ieder systeem binnen het netwerk.
- *Schaalbaarheid*: oplossingen die op kleine, experimentele schaal goed functioneren, moeten ook kunnen werken op grote schaal met hoge informatievolumes en grote aantallen gebruikers.
- *Onderhoudbaarheid* ('sustainability'): oplossingen moeten onderhoudbaar zijn en het moet mogelijk zijn ze verder te ontwikkelen.
- *Personalisatie*: het moet mogelijk zijn om systemen aan te passen aan de behoeften en mogelijkheden van individuele gebruikers.

De resultaten van onderzoek en ontwikkeling voor de innovatie van de informatievoorziening gedurende het afgelopen decennium zijn indrukwekkend en hebben, zoals gezegd, geleid tot een ingrijpende verandering in het functioneren van de informatieketen. Veel van deze programma's zijn evenwel vooral gericht geweest op digitalisering van de traditionele functies en instituties in de informatieketen. Het is de vraag of de daarmee ingeslagen weg voldoende perspectieven biedt voor verdere innovatie. Een aantal programma's, met name die van het Digital Libraries Initiative en ERCIM, nieuwe initiatieven op het gebied van informatieverspreiding via het netwerk, en ontwikkelingen op het terrein van digitale documenten wijzen in een andere, verdergaande richting die wellicht tot een meer fundamentele verandering van de informatieketen zal leiden.

Digitale informatievoorziening: van hybride naar virtueel

De doelstelling van de meeste onderzoeks- en ontwikkelingsinspanningen binnen de genoemde programma's was in het algemeen het ontwikkelen van iets dat (met variaties in de gebruikte terminologie) in het algemeen wordt aangeduid als *digitale informatievoorziening*. Wat we daar precies onder moeten verstaan, is echter niet zonder meer duidelijk. Er zijn talrijke definities in omloop. Een brede definitie zou kunnen zijn: een organisatiewijze van de informatievoorziening waarbij optimaal gebruik wordt gemaakt van de mogelijkheden van de informatie- en communicatietechnologie. Maar beter dan een definitie te zoeken, kunnen we zoeken naar de specifieke kenmerken van de digitale informatievoorziening waarin deze zich onderscheidt van de 'traditionele' informatievoorziening. Die kenmerken zijn de volgende:

- *Digitale informatiebronnen:* de digitale informatievoorziening biedt de gebruiker in ruime mate (zo niet uitsluitend) toegang tot informatiebronnen in digitale vorm. In zijn uiterste vorm leidt dit ertoe dat de gebruiker geen informatie meer op papier verkrijgt.[4]
- *Toegankelijkheid vanaf de werkplek:* de aangeboden informatiebronnen en diensten zijn toegankelijk via het netwerk. De gebruiker raadpleegt de informatie 'op afstand', en hoeft niet meer 'fysiek' naar een bibliotheek- of informatieafdeling te gaan.
- *Gedistribueerde organisatie:* omdat allerlei organisaties die aan de gebruiker informatiediensten aanbieden via het netwerk met elkaar verbonden zijn, ontstaat de mogelijkheid om aan de gebruiker informatiebronnen en -diensten aan te bieden die verzorgd worden vanuit verschillende organisaties op verschillende locaties, maar die de gebruiker ervaart als 'transparant': als afkomstig van één geïntegreerde dienstverlener.
- *Gepersonaliseerde dienstverlening:* door de mogelijkheden om de digitale dienstverlening op allerlei manieren aan te passen aan specifieke wensen van de gebruiker, kan er sprake zijn van een 'persoonlijke' informatievoorziening: iedere gebruiker voelt zich verbinden met een 'digitale bibliotheek' die geheel is afgestemd op zijn of haar behoeften, mogelijkheden, niveau en dergelijke.

Een andere manier om de informatievoorziening te kenmerken, is door te kijken naar twee belangrijke dimensies: de mate waarin gebruik wordt gemaakt van digitale informatiebronnen in relatie tot niet-digitale bronnen, en de mate waarin er nog sprake is van 'fysieke' dienstverleners (bijvoorbeeld gekenmerkt door een specifieke ruimte, één bevoegd gezag, vaste medewerkers, e.d.). We kunnen dan de volgende soorten informatievoorziening onderscheiden (zie ook fig. 1):

- *Traditionele* informatievoorziening: voornamelijk gebaseerd op niet-digitale bronnen, verzorgd vanuit een 'fysieke' organisatie (doorgaans een bibliotheek), nauwelijks toegankelijk via het netwerk.
- *Hybride* informatievoorziening: een gecombineerd aanbod van digitale en niet-digitale bronnen, verzorgd vanuit een 'fysieke' organisatie, maar deels toegankelijk via het netwerk.
- *Virtuele* informatievoorziening: voornamelijk gebaseerd op digitale en gedigitaliseerde[5] bronnen, uitsluitend toegankelijk via het netwerk.

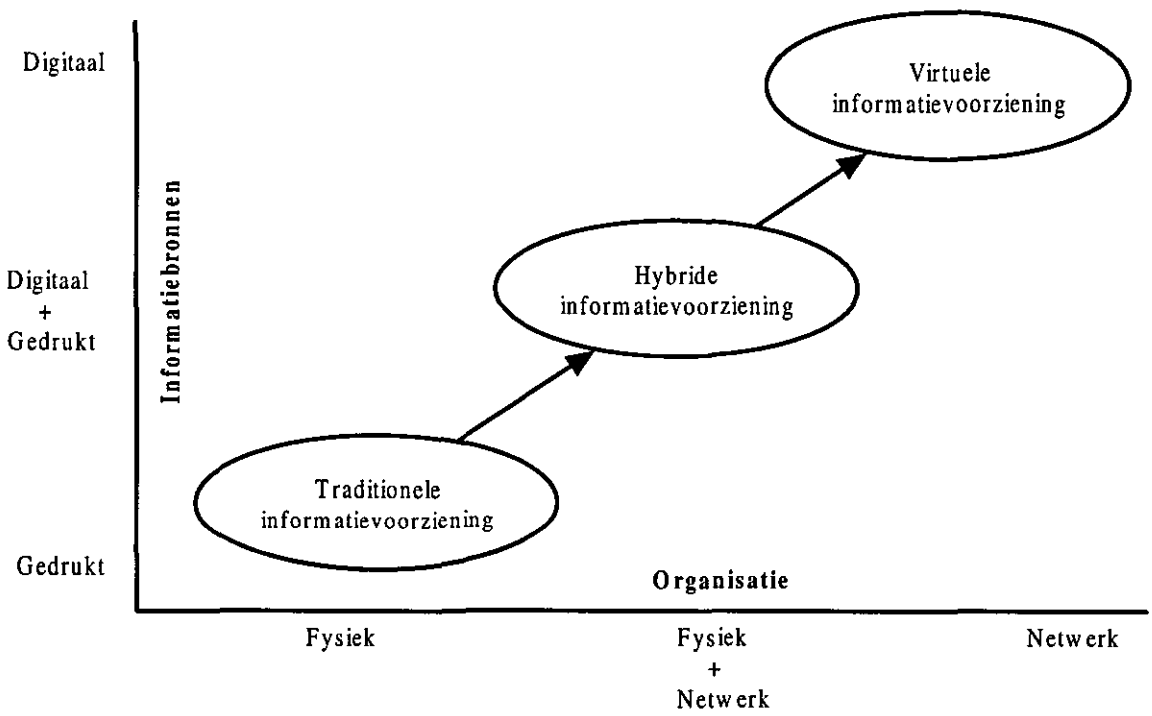


Fig. 1

De ontwikkeling naar hybride informatievoorziening is het duidelijkst waarneembaar bij de wetenschappelijke bibliotheken. Daar zien we een veelheid aan activiteiten die erop gericht zijn om, met de 'traditionele' bibliotheek als vertrekpunt, te komen tot 'hybride' bibliotheken die duidelijk verankerd zijn in de bestaande instituties. Zij zien het als hun taak om zowel de traditionele, niet-digitale collectie als de nieuwe, digitale informatiebronnen te verzorgen.

Deze richting is vooral zichtbaar in Europa waar, in landen zoals Nederland en het Verenigd Koninkrijk en op het niveau van het Kaderprogramma van de Europese Unie, innovatiegelden vooral bedoeld zijn om de bestaande bibliotheekorganisaties tot innovatie aan te zetten. De informatiebronnen die hier aan de orde zijn, zijn tot nog toe in hoge mate beperkt gebleven tot de (digitale versies van) traditionele documentaire vormen zoals boeken en tijdschriften. Om die reden is betrokkenheid van uitgevers bij de ontwikkeling van 'hybride' bibliotheken van relatief groot belang, en spelen daaraan gerelateerde aspecten (zoals auteursrechten, licenties en beschikbaarheid van digitale content) een vrij grote rol.

De ontwikkeling naar virtuele informatievoorziening verloopt heel anders, ondanks het feit dat veel activiteiten op dit terrein worden aangeduid met de term 'digital libraries'. Analyse van projecten en publicaties op dit terrein laat zien men hier werkt op basis van veel abstractere uitgangspunten die geen of vrijwel geen relatie hebben met 'de bibliotheek' zoals wij die tot op heden kennen. Hier vormen primair de gebruiker als doelgroep en het netwerk als technische omgeving de uitgangspunten voor het denken over nieuwe vormen van dienstverlening via het netwerk. De nadruk ligt op geavanceerde technieken voor het creëren en ontsluiten van informatiebronnen en op het ontwikkelen van (veelal nieuwe vormen van) informatie-objecten met 'embedded functionality' waarop binnen het netwerk 'intelligent agents' transacties uitvoeren ten behoeve van de gebruiker. Het einddoel is een geheel nieuwe, geavanceerde informatie-omgeving die op den duur de gehele bestaande informatie-infrastructuur (waarvan publicaties, uitgevers en bibliotheken overigens slechts een deel uitmaken) zou kunnen vervangen. In deze omgeving spelen gedrukte, 'fysieke' informatiebronnen geen rol meer, en is er ook geen sprake meer van voor de gebruiker relevante instituties. De gebruikers (zowel auteurs als eindgebruikers) hebben niet meer te maken met uitgevers en bibliotheken, maar slechts met functies op het netwerk. Daarom duiden we dit hier aan met de term 'virtuele informatievoorziening'.

Hoewel ook bij virtuele informatievoorziening publicaties zeker een rol spelen als informatiebron, zien we binnen deze richting ook veel activiteiten die gericht zijn op andere soorten informatiebronnen (zoals beelddatabanken, gedigitaliseerd historisch archiefmateriaal en dergelijke) en op het ontwikkelen van nieuwe typen informatie-objecten voor specifieke vormen van kennisoverdracht c.q. functies die het verrichten van kenniswerk ondersteunen.[6,7] Zulke activiteiten vinden we vooral in de Verenigde Staten[8], waarbij R&D-projecten doorgaans worden geïnitieerd vanuit de afdelingen 'computer science' van de grotere universiteiten, met een zeer aanzienlijke betrokkenheid vanuit het bedrijfsleven (d.w.z. de IT-industrie). Bibliotheken zijn bij dergelijke projecten niet of slechts zijdelings betrokken. In Europa vinden we deze richting met name in de kring van de European Research Community for Informatics and Mathematics (ERCIM) die een 'Digital Library Initiative' heeft ontplooid.[9]

Een onderzoeksagenda voor de digitale informatievoorziening

In 1998 kwamen verschillende groepen deskundigen uit Europa en de Verenigde Staten in werkgroepen bijeen om te praten over toekomstige samenwerking. In deze werkgroepen hadden voornamelijk deelnemers zitting die betrokken zijn bij projecten van het Digital Libraries Initiative in de Verenigde Staten en bij ERCIM in Europa. Een van de resultaten van deze workshops is een onderzoeksagenda op het terrein van de 'digital libaries'.[10] Dat het concept van de 'digital library' geen betrekking heeft op bibliotheken in de traditionele zin, maar veeleer op nieuwe infrastructuren voor wetenschappelijke informatievoorziening, blijkt wel uit de gepresenteerde voorstellen:

- *Verdergaande samenwerking, bijvoorbeeld tussen groepen die binnen verschillende disciplines werken, maar ook een bredere geografische samenwerking.* Zulke samenwerking wordt mede gezien als een voorwaarde om te komen tot grootschalige c.q. mondiale digitale informatievoorziening waarin de op kleine schaal via meer locale projecten opgedane kennis optimaal benut kan worden. Zulke samenwerking kan op den duur de bestaande organisatorische structuren in hoge mate veranderen.
- *Nieuwe modellen en theorievorming met betrekking tot digitale bibliotheken.* De behoefte hieraan komt onder meer voort uit het feit dat veel R&D-projecten de nadruk leggen op 'development', dat wil zeggen op het ontwikkelen van gebruikersgerichte producten en diensten. Op langere termijn echter kan de innovatie van de informatievoorziening stagneren als er vanuit het meer fundamentele onderzoek te weinig input is. Voorbeelden van belangrijke onderwerpen op dit terrein zijn: de complexe interacties tussen componenten van een wereldwijd gedistribueerde digitale bibliotheek, en onderzoek naar de relatie tussen de werkomgeving en werktaken van de gebruiker en de digitale informatievoorziening. In deze denkwijze wordt de 'digitale bibliotheek' niet meer opgevat als een gedigitaliseerde bibliotheekorganisatie, maar veeleer als een abstracte netwerkinfrastructuur voor distributie en opslag van informatie-objecten.

- *Het ontwikkelen van nieuwe informatie-objecten.* Veel van de 'information resources' die via digitale bibliotheken beschikbaar zijn, zijn digitale versies van gedrukte documenten, of in ieder geval naar het model van het gedrukte document gemodelleerd. Veel onderzoekers zijn ervan overtuigd dat het documentaire paradigma onvoldoende recht doet aan de mogelijkheden van digitale technieken om tot nieuwe, zinvolle vormen van informatieproducten te komen. Hierbij gaat het met name om informatie-objecten met een sterk dynamisch karakter en/of een ingebouwde ('embedded') functionaliteit. [11]
- *Het verder ontwikkelen van de functionaliteit voor het verwerken van zowel nieuwe als traditionele informatie-objecten.* Dit betreft met name functionaliteit met betrekking tot (het benutten van informatie ten dienste van) het uitvoeren van kennisintensieve activiteiten. Hiermee wordt bedoeld dat er binnen het onderzoeksterrein van de digitale bibliotheken ook gekeken wordt naar (softwarematige) voorzieningen op gebruikersniveau voor het creëren en verwerken van informatie die via de netwerkinfrastructuur wordt verspreid.
- *Het ontwikkelen, invoeren en gebruiken van grootschalige, gedistribueerde digitale bibliotheken in de praktijk.* Dit is met name een pleidooi voor een sterke koppeling tussen theorie en praktijk en een oproep om het fundamentele onderzoek daadwerkelijk te benutten door een relatie aan te brengen met de meer toepassingsgerichte ontwikkelingsprojecten. Daarmee wordt aangegeven dat de resultaten van onderzoek op het gebied van digitale bibliotheken niet zonder meer door de bestaande praktijk worden overgenomen, en dat het in feite gaat om het praktisch realiseren van een nieuwe infrastructuur voor wetenschappelijke informatievoorziening.
- *Het opzetten van een algemeen toegankelijke infrastructuur voor het evalueren van toepassingen op het gebied van digitale bibliotheken.* Hiermee wordt erkend dat de meerwaarde van de digitale bibliotheek (in bovenbedoelde zin) niet op voorhand kan worden aangenomen. Een dergelijke innovatie moet geleidelijk worden ingevoerd en voortdurend op zijn merites worden getoetst.

Het is van belang hierbij voor ogen te houden dat deze aanbevelingen afkomstig zijn van die onderzoeksgroepen die zich vooral richten op het concept van de 'virtuele' informatievoorziening zoals we dat hierboven hebben omschreven. Een dergelijk onderzoeksagenda geeft aan in welke richting lange-termijnontwikkelingen gezocht moeten worden, en waar ook voor Nederland op het terrein van meer fundamenteel onderzoek accenten gelegd zouden kunnen worden, en hoe de resultaten van dat onderzoek in praktijk gebracht kunnen worden.

Op weg naar de virtuele informatievoorziening

De virtuele informatievoorziening zoals hierboven omschreven is in feite onderdeel van een nieuwe organisatievorm voor wetenschappelijke communicatie op basis van een mondiale technische infrastructuur. Nadere analyse laat zien dat er hierbij sprake is van drie ontwikkelingen die maken dat de virtuele informatieketen een breuk inluidt met de traditionele informatieketen (en met zijn huidige hybride vorm), en ertoe kunnen leiden dat de toekomstige rol van de daarin opererende instituties onder druk komt te staan. Deze drie ontwikkelingen zijn: verschuiving van de integrale verantwoordelijkheid voor de informatievoorziening naar de wetenschappelijke wereld, integratie van onderzoeksbronnen en -output binnen de informatieketen, en verregaande de-institutionalisering.

Verschuiving van verantwoordelijkheid

In het klassieke model beschouwen wetenschappers en universiteiten zich met name verantwoordelijk voor de productie van wetenschappelijke kennis en de vastlegging daarvan in de vorm van wetenschappelijke documenten zoals boeken, rapporten en artikelen. De verantwoordelijkheid voor de distributie, archivering en beschikbaarheid van die documenten is gedelegeerd aan de overige partijen in de informatieketen zoals uitgevers en bibliotheken. Als gevolg hiervan is ook in Nederland de discussie over de wetenschappelijke informatievoorziening (bijvoorbeeld in het kader van het IWI-platform) voornamelijk beperkt tot de problematiek van het bibliotheekwezen en de (problematische) financiële relatie tussen bibliotheken en uitgevers.

De verschuiving die nu optreedt, houdt in dat wetenschappers en universiteiten ook die aspecten (distributie, archivering en beschikbaarheid) tot hun verantwoordelijkheid gaan rekenen. Het gaat dan om een *integrale* verantwoordelijkheid voor het geheel van de informatievoorziening, dus productie én distributie, archivering en beschikbaarheid van kennis. De wetenschappelijke wereld stelt zich op basis van die verantwoordelijkheid de vraag hoe of men dit geheel wenst in te richten, en ontwikkelt een beleid ten aanzien van genoemde aspecten als kwaliteit, snelheid, functionaliteit, gebruiksgemak en kosten. Daar waar de huidige inrichting van de informatievoorziening onvoldoende aan door de wetenschappelijke wereld geformuleerde wensen voldoet, tracht men daar zelf oplossingen voor te vinden in plaats van die oplossingen te verwachten van derden. Dat laat onverlet dat bij het vinden en implementeren van oplossingen er ook 'externe' partijen kunnen worden ingeschakeld. Waar het om gaat, is dat de wetenschappelijke wereld zich voor het geheel van de wetenschappelijke informatievoorziening verantwoordelijk acht, en de regie daarvan in eigen handen houdt.

Deze trend in de richting van integrale verantwoordelijkheid heeft wellicht voor een deel te maken met onvrede over de mate waarin het gebruik van ICT in de traditionele informatieketen heeft geleid tot verbetering van snelheid, kosten en gebruiksgemak van de informatievoorziening. Maar belangrijker is de groeiende behoefte om informatietechnologie te benutten voor het ontwikkelen van nieuwe informatievormen (zoals de eerder genoemde dynamische, gedistribueerde en interactieve documenten met 'embedded' functionaliteit), voor een vergaande integratie tussen informatievormen (bijvoorbeeld tussen wetenschappelijke verslaglegging, dataverzamelingen en simulaties), en voor nieuwe vormen van computergestuurde visualisatie van onderzoeksresultaten.

Met andere woorden: de wetenschappelijke wereld zoekt naar een rijkere typologie van informatievormen dan binnen de huidige informatieketen (die in hoofdzaak gebaseerd is op traditionele informatievormen als boeken en tijdschriften) beschikbaar is. De behoefte aan zo een verrijking van het repertoire aan informatievormen vloeit voort uit de aard van hoogwaardig wetenschappelijk onderzoek en een besef van nieuwe mogelijkheden die informatie- en communicatietechnologie op langere termijn kan bieden.[12]

Integratie van informatiebronnen
Een belangrijke kenmerk van nieuwe vormen van informatievoorziening die vanuit de wetenschappelijke wereld wordt ontwikkeld, is de toenemende integratie van bronnenmateriaal en onderzoeksinstrumentarium met de verslaggeving van het onderzoek dat op basis van dat materiaal wordt uitgevoerd. Daarmee worden collega-onderzoekers betrokken bij het *onderzoeksproces* zelf, in plaats van bij de *onderzoeksresultaten*. Voorbeelden hiervan zijn de koppeling tussen onderzoekspublicaties en multimediale representaties van onderzoeksmateriaal (datasets, maar ook beeldmateriaal en geluidsopnamen), embedded software en toegang tot simulaties vanuit wetenschappelijke artikelen.

De-institutionalisering
In zekere zin als afgeleide van de meer integrale verantwoordelijkheid van de wetenschappelijke wereld voor de informatievoorziening, ontstaat op verschillende manieren een verschuiving in de richting van de-institutionalisering.[13] In een mondiale netwerkomgeving wordt de wetenschapper minder afhankelijk van institutionele voorzieningen zoals de bibliotheek van de eigen universiteit. Men zoekt informatie daar waar die van de beste kwaliteit is, waar de grootste functionaliteit is, het snelst beschikbaar, en dergelijke. Voor een deel wordt dit al zichtbaar in de trend naar 'domain based services', dat wil zeggen digitale bibliotheken die informatie aanbieden op een specifiek onderwerpsterrein voor een internationale gebruikerskring. Organisatie van de informatievoorziening vindt dan in hoge mate bij de eindgebruiker zelf door personalisatie van de informatievoorziening op de desktop.[14]

Maar de toekomstige virtuele bibliotheek gaat in feite nog verder. Het gaat om een *netwerkinfrastructuur* (conceptueel vergelijkbaar met het Internet), dat functioneert als een disciplinaire werkgemeenschap van wetenschappers die een gemeenschappelijke infrastructuur voor wetenschappelijke informatie-uitwisseling ontwikkelen en benutten. Dergelijke gemeenschappen vormen in het model va de virtuele informatieketen de primaire organisatievorm. Voor zover traditionele bibliotheken en uitgevers daarin een rol spelen zullen, zal dat vooral een ondersteunende rol zijn, in een 'outsourcingrelatie' onder regie van de wetenschappelijke wereld.[15]

Waar het hier met name om gaat is dat er in de toekomst minder deelverantwoordelijkheden zullen zijn die aan afzonderlijke instituties (bibliotheken, uitgevers) zijn toebedeeld en dat de gehele informatieketen onder regie komt van de wetenschappelijke gemeenschap.

Conclusies

Bij de hier geschetste ontwikkeling in de richting van een virtuele informatieketen moet telkens de vraag gesteld worden welke organisatievorm voor de (wetenschappelijke) informatievoorziening op den duur voor de *gebruiker* de meeste toegevoegde waarde oplevert. Die vraag valt niet eenduidig te beantwoorden, omdat allerlei factoren een rol spelen waar we nog onvoldoende inzicht in hebben. Een bijvoorbeeld is het economische model waarop de virtuele informatievoorziening gebaseerd zou moeten worden.

De ontwikkeling naar een meer 'virtuele' informatieketen leidt tot een duidelijke breuk met een organisatiewijze van de wetenschappelijke informatievoorziening die al eeuwen bestaat en heel redelijk voldoet. Actoren op het terrein van de virtuele informatievoorziening verwachten dat die op den duur tot een veel rijkere functionaliteit en informatie-aanbod zal leiden dan innovatie waarbij digitalisering van bestaande documentaire vormen de hoogste ambitie is. Men veronderstelt daarom dat de gebruiker het meest gebaat is bij versterking van de ontwikkeling van de virtuele informatievoorziening, en dus ook bij een verlegging van de innovatie-inspanning van traditionele partijen in de informatieketen naar projecten die gericht zijn op nieuwe functionaliteit en uitbreiding van het palet aan toegankelijke informatiebronnen, met een grote betrokkenheid van en onder leiding van wetenschappelijke gebruikers.

Deze opvatting vinden we impliciet ook terug in het nieuwe onderzoeksprogramma van de Europese Unie, het zogenaamde Vijfde Kaderprogramma. Hierin is het oude *Libraries Programme*[16] verdwenen, en lijkt men van opvatting te zijn dat het niet nodig of zinvol meer is om de innovatie van bibliotheken als zodanig te stimuleren. Min of meer in de plaats daarvan is er nu een programmalijn gekomen op het gebied van 'multimedia content and tools', dat meer aandacht besteedt aan verbreding van het digitale aanbod (met name ook op het gebied van het Europese culturele en wetenschappelijke erfgoed) en aan het ontwikkelen van nieuwe instrumenten voor het creëren, distribueren en benutten van multimediale informatie.[17]

Nederland heeft in veel opzichten een voortrekkersol gespeeld bij het ontwikkelen van de *hybride* bibliotheek, en men heeft laten zien in staat te zijn daar in de praktijk vorm aan te kunnen geven en aldus de wetenschappelijke informatievoorziening in belangrijke mate te kunnen innoveren. Internationaal gezien gaan de ontwikkelingen echter verder, en werkt men in de richting van de *virtuele* informatievoorziening, met mogelijk belangrijke gevolgen voor de inrichting van de informatieketen en de daarin opererende instituties. Het is belangrijk om voor deze ontwikkelingen oog te hebben, en er actief – zij het kritisch – in te participeren. Dat vereist betrokkenheid van alle belanghebbenden: de huidige actoren in de informatieketen, onderzoekers op het gebied van digitale informatievoorziening, en – niet in de laatste plaats – de wetenschappelijke auteurs en eindgebruikers zelf. De uitdaging voor de informatiewetenschap ligt daarbij niet zozeer in het ontwikkelen van nieuwe toepassingen, maar in het verder conceptualiseren van het virtuele model en vooral ook in het bijdragen aan voortdurende en kritische reflectie op de innovatie van de informatievoorziening.

NOTEN

1 Chien, Y.T., 'Digital libraries, knowledge networks, and human-centered information systems'. In: *Proceedings of the International Symposium on Research, Development and Practice in Digital Libraries.* ISDL'97, Tsukuba, Ibaraki, Japan, november 18 - 21, 1997. http://www.dl.ulis.ac.jp/ISDL97/proceedings/ytchien/ytchien.html.

2 Voor een recent model voor een nieuwe inrichting van de wetenschappelijke informatieketen zie: Buck, A.M.; Flagan, R.C.; Coles, B. - Scholar's forum: a new model for scholarly communication, 1999. http://library.caltech.edu/publications/scholarsforum/.

3 Voor een recent overzicht van met name de bibliotheekerichte programma's zie: Brophy, P. - Digital library research overview: final report / Peter Brophy, Centre for Research in Library and Information Management, Department of Information & Communications, Manchester Metropolitan University, August 1999. - www.lic.gov.uk/research/digital/review.rtf.

4 De gebruiker kan uiteraard de informatie wel op papier uitvoeren als dat voor het bestuderen ervan beter is.

5 Onder *digitale* bronnen verstaan we bronnen die vanaf hun creatie/openbaarmaking in digitale vorm beschikbaar zijn. Onder *gedigitaliseerde* bronnen verstaan we bronnen die van oorsprong in niet-digitale vorm beschikbaar zijn, maar die (ten behoeve van gebruik binnen de digitale bibliotheek) in digitale vorm zijn overgezet.

6 Voor een voorbeeld hiervan zie: Phelps, Th.A. - Multivalent documents: anytime, anywhere, any type, every way user-improvable digital documents and systems (Ph.D. dissertation), www.cs.berkeley.edu/~phelps/papers/dissertation-abstract.html.

7 Op basis van concepten als 'networked information objects', 'intelligent agents', 'information transactions' en het marktmechanisme van de informatieketen blijkt er binnen deze richting een nauwe relatie te bestaan tussen digitale bibliotheken en elektronische handel op het Internet. Zie: Adam, N.; Yesha, Y. e.a. - Strategic directions in electronic commerce and digital libraries: towards a digital agora. - In: ACM computing surveys, 25(1996)4, p. 818-835.

8 Bijvoorbeeld binnen het DLI-programma met topics als: Information object models, User interfaces and human-computer interaction, Information discovery, Archiving, Interoperability, Authentication and security, Semantic interoperability, Scaling, en Economic, social and legal issues. Inmiddels is Pahse 2 van het DLI-programma van start gegaan. Zie: Digital libraries initiative - phase 2, www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm.

9 ERCIM (www.ercim.com) heeft zich op dit terrein vooral geprofileerd door middel van de zgn DELOS-workshops. Onderwerpen die hier aan de orde komen hebben vooral betrekking op zaken als intelligente retrieval en filtering, gebruikersinterfaces, representatie, visualisatie, meertalige systemen, e.d. Zie: www.iei.pi.cnr.it/DELOS//.

10 Schäuble, P.; Smeaton, A.F. - A Research Agenda for Digital Libraries / Joint NSF-EU Working Group on Future Directions for Digital Libraries Research, October 12, 1998. Zie ook www.dli2.nsf.gov/workgroups.html.

11 Hiermee worden ondermeer bedoeld documenten waarvan de inhoud in real-time, bij het raadplegen, uit andere informatie-objecten wordt samengesteld, en documenten die 'ingebouwde' programmatuur bevatten waarmee bewerkingen op de inhoud mogelijk zijn.

12 Hiermee hangt samen de afnemende betekenis van het traditionele concept van het 'document' als basis voor de organisatie van de informatievoorziening. Zie: Mackenzie Owen, J.S. - Het document aan het einde van de 20ste eeuw. – Voordracht op de 8e Dag van het document, Ede, 14 september 1999. http://www.org.uva.nl/bai/home/jmackenzie/pubs/jmo-dvd.htm.

13 In het verlengde hiervan is er ook sprake van verdergaande *internationalisering*. Dit houdt in dat niet alleen de wijze waarop de informatievoorziening binnen een bepaalde instelling wordt georganiseerd, maar ook de manier waarop dat binnen een bepaald land gebeurt, van afnemend belang is. De virtuele bibliotheek kent intrinsiek geen institutionele of nationale grenzen.

14 Deze vorm van personalisatie, waarbij de gebruiker verschillende informatiebronnen en -leveranciers integreert binnen zijn eigen werkplek, wijkt fundamenteel af van een gepersonaliseerde dienstverlening waarbij één institutie (bijvoorbeeld een bibliotheek) op basis van gebruikersprofielen aan individuele gebruikers diensten op maat levert.

15 Voor een bespreking van de relatie tussen bibliotheken en uitgevers vanuit het perspectief van de informatiemarkt, zie: Odlyzko, A. - Competition and cooperation: libraries and publishers in the transition to electronic scholarly journals. – In: Journal of Electronic Publishing, 4(1999)4. www.press.umich.edu/jep/04-04/odlyzko0404.html

16 Het vierde Kaderprogramma kende nog een 'Libraries Programme', met als actielijnen: de ontwikkeling van netwerkmogelijkheden voor interne bibliotheeksystemen, de koppeling van bibliotheken onderling, en de bibliotheek als intermediair tussen de gebruiker en het netwerk.

17 Zie http://www.echo.lu/libraries/en/libraries.html en http://www.echo.lu/digicult/

Document information standards and longevity
XML will not solve the problem of longevity

K. van der Meer (1) and J.J.M. Uijlenbroek (2)
(1) Delft University of Technology
Dept. of Information Systems and Software Engineering
Zuidplantsoen 4
PO box 356
2600 AJ Delft
The Netherlands
(2) Het Expertise Centrum
Jan Willem Frisolaan 3
2517 JS 's-Gravenhage
The Netherlands

Abstract
Standards like SGML, XML and PDF are used when electronic documents are archived; they play a role in document longevity. But standards have a restricted life span. The everlasting standard is an illusion; at the end of their lifecycle standards will fade away. Even XML will become obsolete in the future. To substantiate the evolution of standards currently existing standards are summarized. This paper presents a strategy toward the digital document longevity problem based on the evolution of standards.

Introduction
It may come as a surprise that longevity of document information systems (including digital documents) could be a problem at all. It is so easy to move, copy and cut and paste digital document information. One can so easily construct a new document by using available document parts or merge parts of document databases for the purpose of publishing. It is a conceptual oxymoron that document information seems to taint and wither under one's hands.

Nevertheless, the digital longevity problem is a real problem, as was noticed by people needing to access old digital documents. The problem was recognised several years ago. The Dutch Bureau of Digital Longevity was founded in 1991, the book 'Preserving the Present', drawing a lot of attention at that time, was published in 1993 ([Biks93]) and Rothenberg's well known article appeared in the Scientific American in 1995 ([Roth95]). In the last few years many organisations needed to convert digital documents or had even to face their loss. Costly conversions and document loss drew attention to the longevity problem. A description of the growth path in The Netherlands is given in [Meer98].

In spite of all that, still today many people find it difficult to appraise the relevance of the questions. Proposals of answers are even more difficult to appraise. The basic problem is: if we want to guarantee that a digital document (containing content, and sufficient content meta data and formal meta data to make its function understood) will be preserved for 50 years, how can that be attained? On what data carrier can it be preserved? If it is migrated to some other carrier and some other platform, can its provenance still be proved? What will its legal status be?

How to recognise digital annotations, added later, to it? On a higher level of thinking, what is the relation between the architecture and the development of a policy and strategies to preserve digital documents? Could an archive issue a Service Level Agreement (SLA) to promise to society that documents can still be accessed and serviced in 50 years' time?

Early 1999, a discussion erupted on the best strategy for preservation. Rothenberg ([Roth99]) pointed out advantages of emulation; Bearman ([Bear99]) argued against this and advocated migration. The dispute as such helped to draw attention to the longevity problem, too. In their dispute, the basic fact that any simple, universally applicable, one-time fix strategy (if that exists) will be dependent on what to preserve, is not mentioned.

Standards are vital

Organisations exchange lots of documents with their environments. The sheer reason of the existence of many documents is their serving as a means of communication between organisations or units. Records witness transactions between organisations, stem from CSCW, or originate from outside the organisation that keeps the records. Document sharing literally depends on standardisation. Collaboration of organisations may literally depend on the architecture of digital document information systems.

Standards are an architectural element. Although it is often recognised that it is useful to standardise building blocks of information systems and documents (e.g. [CCSD98], [Cerf97], [Meer98], [Möll99] and [Tomb97]), the importance of standards for the goal of longevity has often been overlooked or disputed ([Bear99], [Heds97], [Hors97], [Mosc97], [Roth99] and [Shea97]).
Firstly, it has been stated that standards or any architectural element cannot solve the problem of digital longevity. Still, as pointed out, this does not mean that standards cannot be an essential part of the solution. Secondly, it has been stated that even architecture is subject to change. As Bearman puts it: 'No computer technical standards have yet shown any likelihood of lasting forever – indeed most have become completely obsolete within a couple of software generations. This fallacy currently feeds the imaginations of those proposing a Universal Preservation Format who are unfortunately not alone in imagining their "standards" will indeed last forever' ([Bear99]). There is a point. Standards do not live forever. Technical developments and improvements continue to lead to new possibilities and to evolution of standards. But this cannot be used as a reason to dismiss all standardisation for the purpose of longevity!

Standards help to structure the objects of our discussion. They thus play a role in the preservation issue. The architecture of document information systems, including the functionality and including the standards on which they are based, must answer longevity requirements, otherwise it is void. This leads to the question how document information systems, containing digital documents and their standards can be carried through time. For a preliminary answer to that question, let us browse through some standards. We limit ourselves to standards on (provision of) document information and comment on the longevity aspect.

XML: today's most popular standard

The acceptance of the eXtensible Markup Language (XML) standard, version 1.0, in 1998 by the World Wide Web Consortium (W3C) has been of enormous importance for digital document provision. Many organisations have realised the importance of XML. Microsoft, Netscape and other major organisations in this field offer XML support.

Many websites for e-commerce are XML-ed. A huge amount of effort is spent to comply with XML and with the standards related to it. Examples of emerging and developing standards related to XML are XSL (eXtended Style Language), Xpointer, XHTML (eXtensible Hypertext Markup Language), OFX (Open Financial eXchange), CDF (Channel Definition Format), RDF (Resource Description Framework), P3P (Platform for Privacy Preferences) and others. There are some 20 standards related to XML. Moreover, new DTDs, to be used with XML, seem to be published on the Internet every week. In line with the growing popularity of XML as compared to SGML, XML is often presented as a major step forward for the digital longevity problem.

Unfortunately it is not the first effort to realise a standard for structuring documents. As long ago as 1986 the International Organization for Standardization, ISO, accepted two standards, both with this purpose: Office Document Architecture (ODA) and Office Document Interchange Format (ODIF), ISO 8613, as well as Standard Generalized Markup Language (SGML), ISO 8879.

The SGML architecture has surely contributed to the concept of a document. SGML changed the image of a document from a physical reality into a processing metaphor for many different information-bearing forms, such as e-mail, journal articles, bulletin notes, manuals, multimedia documents, and from a page layout-based (human-readable) object to a content-oriented (computer-reusable) object. SGML is not a simple and cheap mechanism. SGML proved to be complex and difficult to use in practice.

Its use demands a lot of effort and money. Broad usage of SGML has been limited to two sectors: publishers and manufactories of product manuals, in other sectors SGML was used less generally.
So the user group of that important development was limited.

The fate of ODA/ODIF was even worse. This standard is even more extended than SGML. It was suspected and found to be very expensive to implement and use. Although it was an official ISO standard, the use of ODA never really took off.

Document standards

Despite XML's current popularity, it is just one of many document standards. Present document standards are listed in the appendix. Document standards are categorised as follows:

- document structuring of which SGML is an important example;
- printer language standards, for example the well known PDF;
- interoperability standards which define the way access is given to documents in a distributed environment;
- meta data elements standards (e.g. the Dublin core);
- content access for information retrieval;
- information content structures;
- character sets (e.g. ASCII).

There are also a number of standards that do not fit into these categories. These are presented in the category 'others'.

"One of the great things about standards is that there are so many different ones to choose from" is a common good-humoured quote. It proves to be wholly unjustified! Admittedly many standards exist. But when it comes to the stability of the architecture and especially the stability of the standards of document information systems and the digital documents, their use is limited.
Document information systems, like in Information and Communication Technology (ICT) systems in general, develop very fast. Both the architecture of the system functions and the architecture of the documents are subject to change upon time. The speed of ICT developments tends to draw attention away from the stability of architecture with time. As a consequence, there are only few standards on functions and documents for which we would dare to advise positively on the issue of longevity. A Service Level Agreement (SLA) for digital documents should better not be issued at this moment.

Taking a closer look at the standards in the Appendix the evolution of standards becomes clear. Open Document Architecture (ODA) and Open Document Interchange Format (ODIF) is a striking example to illustrate the life cycle of standards. It was a lot of work to describe ODA/ODIF (it is extended) and a lot of scientific work was based on it. The standard, accepted by ISO in 1986, is currently fading away.

Notions concerning the evaluation of standards

In order to evaluate families of standards from the digital longevity point of view, the following aspects are distinguished concerning document information systems:

- Software functions of the document information systems: storage, (remote) accessibility of documents, interoperability, retrieval aspects;
- Documents have a content, which is the dominant aspect in standards that enable document structuring;
- Documents have a description, on which meta data standards and the thesaurus structures focus.

Authenticity is of vital interest for archival purposes. It includes provenance and requires longevity, so that future generations can access documents including meta data in their original formats. This implies that not only the documents have to be stored, but also their complete environment, hardware and software. The cost of such a set-up is very high. Sooner or later some kind of information loss has to be accepted. Consequently, it is necessary to determine to what extent a stored document should correspond to its original. One should not freeze all document information systems in order to guarantee digital longevity, but rather enable future computer-scientists-and-historians to reconstruct documents and the information therein. This idea corresponds to the way historians work today: reconstructing documents and information from the past.

At first glance, if one abandons the demand regarding the original format of a document but insists on reusability, one ends up with standards like the SGML/XML family. If the demand regarding reusability is dropped the PDF standard satisfies. Sometimes publishers use both kinds of standards.

Organisations that recognised the advantages of logically structure documents have spent fortunes to convert documents into SGML format. From 1986 until 1996 no alternative to SGML existed. Rothenberg states: 'In fact, if SGML had been adopted as a common interlingua (a commonly translatable intermediate form) among word processing programs, it would have greatly relieved the daily conversion problems that plague most computer users. This however has not occurred, implying that even well designed standards do not necessarily sweep the marketplace.

Nevertheless, converting digital documents into standard forms, and migrating to new standards if necessary, may be a useful approach and is probably the only approach if one assumes that a true long-term solution (e.g. an everlasting standard) will never be developed. This suggest that standards may play a minor role in a long-term solution by providing a way to keep meta data and annotations readable.' ([Roth99]).
Still, there are organisations that have to write off part of the investments in SGML developments. This is a painful fact that we want to avoid for XML documents.

The architecture must answer longevity requirements. In the case of document structures and printer languages, the less complex a standard is, the lower the cost may be to convert a digital document from one application to another. The consequence is that more information will be lost on storage of documents. There is a trade-off between the current possibilities (cost) of preventing loss of information and the necessary expertise (cost) of reconstructing a document and the information therein. Organisations aiming at the preservation of historical artefacts are familiar with this trade-off: archives routinely appraise documents to decide which ones will be preserved, because of their organisational, cultural or evidential value, and reconstruct information in their historical context. This discussion results in a new question: which elements of SGML/XML are necessary for longevity?

So far the evaluation focussed on the documents content and possible loss of information.
This line of reasoning is to be applied to the document description, the meta data, too. Meta data describe the context of a document necessary to understand and interpret it. Partial loss of meta data does not by definition makes a document obsolete. Compare it to the loss of information of paper documents from the past, where lots of meta data were lost and have been reconstructed by historians. Again a trade off arises: what are the cost of adding relevant meta data compared to future recovering of lost meta data and what document elements are so important that loss of meta data cannot be allowed.
A certain knowledge area structured in a thesaurus is by definition cultural and time based. Today's thesaurus may be of importance to future generations concerning the way a certain knowledge area was structured, but will probably not satisfy future demands concerning accessibility. It is possible to draw up lean meta data models for digital longevity?

The same question must be raised concerning systems functionality. It may be necessary to preserve certain system functions. An example is the interoperability issue. What functionality of today do we want to preserve?

Ensuring digital longevity: architectural approach and standards to system design

There is no standard, which ensures digital document longevity. We argue that as long as technology evolves standards will become obsolete. However, the use of standards is important for a number of reasons. If an architectural approach to system design is used, and it is implemented using component based development technologies, one may expect that standardised, for instance XML-based, components can rather easily be substituted. Consequently, digital longevity influences information system design (e.g. the architecture and component based) and information system maintenance. Furthermore, it implies the choice and use of standards and the replacement of components based on certain standards. The basic strategy in system maintenance should be reducing the frequency of converting documents from one standard to another and reducing the cost and information lost in converting. Consequently, the lifetime of standards should be extended and converting from one standard to the other (its successor) should be made as easy as possible.

Conclusion

One day, XML will be a bit obsolete, like SGML is being made a bit obsolete these days. No standard for digital document longevity may survive eternally. Still, standards are needed and will continue to be used to structure document information systems and digital documents.

Standards must face longevity; otherwise they are void. A survey gives rise to the presumption that no current standard can be recommended for longevity purposes. However, standards are necessary to enable preservation. They must be developed starting from existing standards, such as SGML/XML and other families of standards. They should be based upon wanted behaviour and foreseen requirements of current digital documents in the future. Backward compatibility of new versions of standards must be zealously defended.
Work is in progress to construct these standards.

References

N.B.: In accordance to a guideline in ISO 690-2 the dates of consult of electronic sources have been added.

[Bear99] D. Bearman: Reality and chimeras in the preservation of electronic records. D-Lib Magazine 5(4), (1999), April 1999. [May 3, 1999].

[Biks93] T.K. Bikson and E.J. Frinking: Preserving the Present / Het Heden Onthouden. SDU Publishers, 's-Gravenhage, 1993.

[Cerf97] P. le Cerf, L. de Bremme and R. Schockaert: Standards for electronic document management. In: Proceedings of the DLM-forum, Brussels, December 1996. Luxembourg: Office for Official Publications of the EC, 1997. Pp. 217-222.

[CCSD] Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System. CCSDS 650.0-W-5.0 White book. April 21, 1999.

[DMA] http://www.aiim.org/dma. [September 15, 1999].

[Dubl] http://purl.oclc.org/metadata/dublin core. [May 10, 1999]

[ECHO] http://www2.echo.lu/oii/en/archives.html. [September 15, 1999].

[Heds97] M. Hedstrom: Research issues in migration and long-term preservation. Archives and Museum Informatics 11 (3-4), (1997), 287-291.

[Hors97] P. Horsman: Digital Longevity: policies on electronic records in the Netherlands. Archives and Museum Informatics 11, (3-4), (1997), 235-240.

[ISO46.4] http://www.iso.ch/liste/TC46SC9.html. [September 15, 1999].

[ISO46.9] http://www.iso.ch/liste/TC46SC4.html. [September 15, 1999].

[Kasd98] B. Kasdorf: SGML and PDF – Why we need both. J. Electronic Pub. 3(4), (1998). [June 11, 1999].

[Lawr99] S. Lawrence and C.L. Giles: Accessibility of information on the web. Nature 400 (1999), 107-109.

[Marc97] Y. Marcoux and M. Svigny: Why SGML? Why now? J. Am. Soc. Info. Sc. 48(7), (1997), 584-592.

[Möll99] J. Möller, J. Boogaarts and H. Nijborg: Minimum functionele eisen voor elektronisch archiefbeheer volgens US DoD 5012.2-STD (Minimum functional requirements for electronic records management according to US DoD 5012.2-STD). Archiefbeheer, artikel 5750, 1999 (in Dutch).

[Meer98] K. van der Meer: Documentaire informatiesystemen (Document information systems). 3e gewijzigde druk. NBLC, 's-Gravenhage. 1998. (In Dutch).

[Meer99] K. van der Meer and H.G. Sol: The design of document information systems. In: A. Kent (executive editor): Encyclopedia of Library and Information Science. Marcel Dekker Inc., New York [etc.], Vol. 64, 1999. Pp. 68-95.

[Mosc97] L. Moscato: Australian approaches to policy development and resulting research issues. Archives and Museum Informatics 11(3-4), (1997), 241-250.

[ODMA] http://www.activedoc.com/odma. [September 15, 1999].

[Over95] P. Over, W.E. Moen, R. Denenberg and L. Stovel: Z39.50 Implementation experiences. NIST Special Publication 500-229, US Department of Commerce, NIST, Gaithersburg, 1995.

[PDF] http://www.purepdf.com. [September 17, 1999].

[Roth95] J. Rothenberg: Ensuring the longevity of digital documents. Scientific American 272 (1), (1995), 24-29.

[Roth99] J. Rothenberg: Avoiding technological quicksand: finding a viable technical foundation for digital preservation. A report to the Council of Library & Information Resources (CLIR), January 1999. (http://www.clir.org/pubs/reports/rothenberg/pub77.pdf. [April 19, 1999].

[Shea97] G. O'Shea: Research issues in Australian approaches to policy development. Archives and Museum Informatics 11 (3-4), (1997), 251-257.

[Tomb97] K. Tombs: Governmental, industry and user perspectives of achieving standard storage mechanisms for long-term archival activities. In: Proceedings of the DLM-forum, Brussels, December 1996. Luxembourg: Office for Official Publications of the EC, 1997. Pp. 210-216.

[Whit98] E.J. Whitehead: Collaborative authoring on the Web: Introducing WebDAV. Bulletin of the Am. Soc. Info. Sc., Oct/Nov 1998, 25-29.

Appendix. Contemporary document standards

1. Document structuring.

- Standard Generalised Mark-up Language, SGML (ISO 8879) to describe the logical document structure. SGML is well known, many books on it are available; Marcoux and Svigny offer a gentle introduction ([Marc97]). It requires the use of a Document Type Definition (DTD). An SGML document is marked up with 'tags' that have been specified by the DTD.
- Examples of DTD's have been published in ISO 12083, together with an SGML declaration that restricts the original freedom. The ISO 12083 DTD's are meant to be both user's guides and building blocks. ISO 12083 provides architecture to facilitate the creation of DTD's for a specific goal.
- Document Style Specification and Semantics Language, DSSSL (ISO 10179), the partner for SGML to describe the document layout structure.
- HyperText Mark-up Language, HTML for the mark-up and hyperlink facilities of Internet documents. Four subsequent versions of HTML have officially been agreed upon. HTML 4.0 is being standardised as ISO Draft International Standard (DIS) 15445. It is assumed that no HTML 5.0 will ever be needed, due to the rise of XML. HTML 'is' a Document Type Definition (DTD) in line with the SGML framework. HTML offers some possibilities to structure documents in a logical way and is used for webdocuments.
- Cascading Style Sheets, CSS (no ISO standard), viz. the DSSSL for HTML.
- eXtensible Mark-up Language, XML. XML is not (yet) an ISO standard, but issued as a recommendation by the World Wide Web Consortium (W3C). XML is less complicated than SGML and therefore easier to use. Under XML it is possible to bypass the use of a DTD. It is allowed to have XML documents well formed, i.e. if elements nest properly and could be added to an existing DTD, the document still can be validated.
- XHTML (eXtensible Hypertext Markup Language), a reformulation of HTML 4.0 in XML 1.0, accepted by W3C.
- The standards related to XML of which working drafts exist or that are being developed, they were pointed at in the introduction.
- HyTime is a kind of SGML for Hypermedia and Time-based documents (ISO 10744). It enables to mark up textual information in digital format, graphics, audio and video; it is an extension in scope of SGML.
- Standards for structuring multimedia documents. Synchronised Multimedia Integration Language (SMIL) is an HTML-like mark-up language to synchronise text, graphics, audio and video is an example.

The 'other' standard for document structuring mentioned in the introduction is ODA/ODIF (Office Document Architecture / Interchange Format, ISO 8613). Conceptually it is equivalent to SGML and DSSSL together. It did not bring much from the view of digital longevity. As indicated in the introduction, the ODA/ODIF standard was hardly used. It came too early, if it would have been viable at all.

2. Printer language standards.

Portable Document Format, PDF ([PDF]), can be seen as the printer language standard. As Kasdorf explains, it is something else than the mark-up language structures ([Kasd98]). It is based upon PostScript. PDF is not an official standard, but is proprietary to Adobe®. PDF files are fixed digital images of the printed page, with the layout exactly reproduced.The advantages of PDF are the ease of production (in fact a by-product of PostScript), the fidelity to the printed book or journal, and the simplicity to generate printed output from electronic files.

PDF is a de facto standard. The ISO printer language standard is Standard Page Description Language (SPDL), ISO 10180. SPDL does not seem to have enough advantages over PDF. While SPDL is hardly supported in practice, PDF is ubiquitous. Common document management software (Documentum, DOCS Open, OpenText, Saros, Seasoft) can manage PDF documents. Nowadays, everyone with a PC seems to have a (free) Adobe® Acrobat reader. Fall 1999, the Acrobat reader software to read and print PDF seems to have been downloaded over 20 million times. But more software is necessary to use all features of PDF, among others Acrobat Capture (not free).
The more recent versions of the Acrobat reader, of which version 4.0 (March 1999) is the last one, are backward compatible to former versions. Documents tagged in a recent format are not readable in an old version. The life expectancy of this standard, PDF, may be dependent on its installed base. Its longevity may be proportional to the longevity of the owner organisation.

3. Interoperability of documents.

- ODMA (Open Document Management API, [ODMA]) and
- DMA (Document Management Alliance, [DMA]).
  Both these are standards aiming at interoperability: access to documents in a distributed environment, open to different document management software and different applications. Of ODMA, a one-to-many solution, version 2.0 is to a certain extent accepted in the market. Many vendors advertised that their document management software is 'ODMA compliant'. Of DMA, the many-to-many option in which software of different vendors should be capable to interoperate, the acceptance of version 1.0 seems to lag behind. ODMA and DMA are owned by AIIM.
- WebDAV (World Wide Web Distributed Authoring and Versioning, [Whit98]) is a distributed authoring protocol for Internet documents on which much work has been done since 1997. WebDAV is 'owned' by the World Wide Web Consortium, W3C. It is at this moment not well known. It is not yet much used, but it offers prospects for organisations with Internet and Intranet infrastructures.

It is not clear whether any one of the standards for document interoperability is stable and whether it will receive sufficient mass. Support for these standards in the future may be a problem. Longevity aspects seem to have been neglected until now.

4. Meta data elements

A lot of work has been done on the standardisation of meta data for digital documents since about 1970. Examples of sets are the following.

- MARC set,
- ISAD (G) set,
- Dublin core.

The Machine Readable Cataloging (MARC) set and its 'user's guide', the Anglo-American Cataloguing Rules (AACR2) are well known in the library. Still, minor differences occur in the use of cataloguing rules between libraries. Most people feel that MARC is not suited for archival record management. A similar tool for archives and record management is developing.
UNESCO set up the International Standard for Archival Description (General) (ISAD (G)) in 1994. Alas, despite the high patronage there is still only a very limited installed base of ISAD (G).
The Dublin core set ([Dubl]) is set up for digital library materials. It is wider applicable. The use of it has not yet crystallised out. It is not used overmuch. Lawrence and Giles state that 0.3% of the websites containing meta data use the Dublin core standard ([Lawr99]). The standardisation process in this area is converging only very slowly.

5. Content access for information retrieval

Content access will probably be defined by ISO 23950. This is at the moment a Draft International Standard (DIS). ISO 23950 supersedes the well known American National Standards Institute (ANSI) Z39.50 information retrieval protocol and related ANSI standards, as well as some contents of the ISO standards in the range of ISO 10160 up to and including ISO 10166. It sees to the interaction protocol in a retrieval process between the client and the server and to the query language.
The Z39.50 standard is widely known ([Over95]) and although there were apparently differences in its implementation, its content and that of the standards ISO 10160 - 10166 seem to have been respected. The ISO 23950 standard could be a basis for a long-time standard, although there is only hope and not yet proof for that.

6. Information content structures.

The information content structures of library materials have been standardised by thesauruses. There are two types of these: the monolingual thesaurus, ISO 2788, and the multilingual, ISO 5964. ISO 2788 is generally known and many document information packages claim to be conforming to that standard. ISO 5964 is less often used.

7. Character sets

The list of standardised character sets contains ASCII, the 7-bit American Standard Code for Information Interchange (ISO 646), ISO 8859 (the Latin series), ISO 6937 (Teletext and Videotex) and finally Unicode (ISO 10646). ISO 8859 and ISO 6937 include ISO 646; ISO 10646 includes the other standards. Apparently, there is backward compatibility to the older standards (they are still used). Unicode is regarded as a very stable standard. Its lifetime may be centuries rather than decades, which is the case for most standards. Or will a new character set including the smiley ☺ and its successors be standardised?

8. Others

A kind of catalogue for document information standards that are relevant from the viewpoint of document information provision but not for ICT can be found at [ISO46.4], [ISO 46.9] and [ECHO].
Examples of ICT standards that are less important for document information provision are

- http and ftp etc.,
- CORBA,
- COM and DCOM,
- OLE,
- DLL,
- The ISO 9660 'High Sierra' file system for CD-ROM,
- The ISO 11560 file system for 5¼ inch / 130 mm WORM (Write Once, Read Multiple) disks,
- The Workflow Management Coalition API standards.

Query by Navigation on the WWW

B.C.M. Wondergem, M. van Uden, P. van Bommel, and Th.P. van der Weide
Computing Science Institute, University of Nijmegen
Toernooiveld 1, NL-6525 ED, Nijmegen, The Netherlands
Tel: +31 24 3653147, Fax: +31 24 3553450
E-mail: bernd@cs.kun.nl

Keywords: Information Retrieval, Hypertext, Query by Navigation, World Wide Web, Index Expressions.

Abstract

Searching information from a large and dynamic information space causes several problems, concerning, for instance, dynamic and vague information needs, too broad queries, and correctness and sensibility of descriptors. These problems may be attacked by navigational query formulation strategies which are available for stratified architectures. However, stratified architectures cannot be easily constructed for large and dynamic information spaces. In this article, we show how navigational query formulation and exploration can be employed on the WWW by using linguistic (as opposed to statistical) refinements. Grounded in the theory of navigational networks for index expressions, we introduce our tool, the INdex Navigator (INN), for searching and navigating the WWW. The INN is a dynamic electronic service system for the WWW.

1. Introduction

Searching information from a large and dynamic information space causes several serious difficulties. A number of these concern query formulation. According to [3,11], a major problem is (caused by) the inherent vagueness of information needs. Therefore, formulating the information need concisely, without an explicit description of the expected result, is very difficult. In order to increase the user's knowledge about the field of interest, IR systems should enable users to explore topics of interest. Related to an increase in knowledge are shifts in interests ([3]). Retrieval systems should thus support interactive reformulation. A third major problem concerns constructing (syntactically) correct and (semantically) sensible complex descriptors ([9,10]). Letting the user select descriptors from a set of (correct) options bypasses this problem. Fourth, broad queries often result in low precision. IR systems should thus aim at preventing imprecise queries.

Formulating queries by navigation in an abstraction of the information space eases the mentioned problems ([8,5,14]). To this end, stratified architectures have been developed containing an ancillary layer that forms an abstract description of the contents of the information space ([1,5]). This meta layer can depict an overview of the concepts present. This helps users in exploring their field of interest. Searchers can then formulate their need by recognizing rather than formulating relevant concepts. The vagueness in the information need can be further decreased by concept exploration: by inspecting actual documents that correspond to a concept. In this way, the user can learn what the concept means. This is the second way in which navigation aids exploration.
Since the IR system generates the overview, it can guarantee correctness and sensibility of the offered descriptors. For instance, the descriptors can be taken from available documents.
Shifts in interest are naturally supported by selecting a different direction during navigation in the overview. Finally, navigational formulation techniques enable users to iteratively select more specific descriptors. In general, this eases the way to descriptors of proper specificity. Concluding, we advocate the combination of searching and exploration based on navigation in an ancillary structure. This integration of hypermedia and information retrieval is also advocated by, for instance, Waterworth & Chignell ([13]) and Lucarella & Zanzi ([8]).
However, the size and dynamics of certain information spaces imply that a complete and up to date abstraction cannot be constructed. Consequently, navigational query formulation in ancillary layers is not directly applicable to these information spaces.

In this article, we show how a very large and highly dynamic information space, the World Wide Web, can be abstracted and navigated. We developed the INdex Navigator (INN), a dynamic information system for query formulation on and exploration of the WWW. The INN is based on Query by Navigation (QBN). QBN ([5]) is a navigational way of query formulation in a stratified architecture based on index expressions ([6]). Navigational networks for index expressions allow navigation over linguistically motivated subexpression links. Although the INN is developed for the WWW, our approach is mutatis mutandis applicable to all (dynamic or static) information spaces. The required changes only involve the communication of the INN system with the search facilities used to access the information space. The INN system is available from
http://www.cs.kun.nl/is/inn

This article has the following structure. Section 2 explains the construction of a stratified architecture based on index expressions. Section 3 explains QBN. Insight in the construction of the INN system is given is section 4. Its workings are illustrated by an example session. We provide ways for further research in section 5.

2. Stratified Architecture for Index Expressions

2.1 Search Support using Layers

The stratified architecture augments a set of documents with an ancillary structure, called the hyperindex. This hyperindex forms a more abstract description of the contents of the documents. It provides a conceptual overview of the information carried in the documents. The special form of our hyperindex, which is explained in the next section, allows structural navigation. Concept exploration is supported by transferring between the hyperindex and the actual documents that correspond to a concept in the hyperindex.

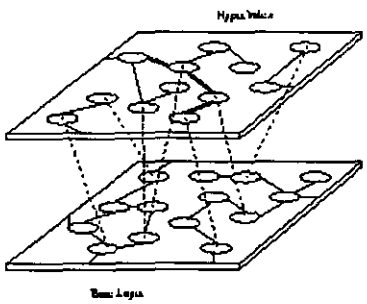


Figure 1: Stratified Architecture

The stratified architecture, as depicted in figure 1, consists of two layers, the base layer and the hyperindex, which are connected through the beam relation. The base layer contains the available documents. These documents may be linked, for example by hyperlinks. By traversing these links, as is usual in WWW context, navigation in the base layer may take place. We will not study this further in this article. Each document is indexed, or characterised, which results in a set of descriptors, a characterisation, per document. These descriptors reside in the hyperindex.

The beam relation connects the base layer with the hyperindex. That is, the beam relation connects documents in the base layer with their characterisation. In general, the links between both layers are be bidirectional, allowing traversal in both directions. From a document, the user may transfer himself to one of the descriptors of the characterisation (beam up). From a descriptor in the hyperindex, the user can perform a beam down, which transfers him to documents that are about the descriptor.

The hyperindex forms an overview of the documents based on their characterisations. If index expressions are used for characterisations, a hyperindex with a fine-grained structure is constructed. This structure is used in step-wise navigation.

2.2 Index Expressions for Document Characterisation

In our case, the descriptors that form document characterisations are index expressions. In the hyperindex, index expressions are connected with their subexpressions and superexpressions

Index expressions are built from terms and connectors by structural composition. Terms denote keywords, concept names, adjectives, gerunds, and attribute values. Connectors denote relations between terms in the form of prepositions (showing place, position, time, or method) and some present participles (e.g. using, having, and being). Structural composition, using the structural operator add, gives index expressions their (tree-like) structure.

Index Expressions
For given sets of terms T and connectors C, index expressions are defined by the following two cases (Terms) Each term t is an index expression. (Composition) If I and J are index expressions and c is a connector, then add(I,c,J) is an index expression.

The composed index expression add(I,c,J) is obtained by adding subexpression J to I via connector c. Using nesting of the structural operator, index expressions can be made more specific, by deepening the tree structure, or more general, by broadening the tree structure. This is illustrated below.

Consider as terms conference, information, and retrieval. Furthermore, consider as connectors on and in. The phrase conference on retrieval in Amsterdam is an index expression. It is denoted as add(add(conference,on,retrieval),in,Amsterdam). As another example index expression, consider add(conference,on,add(retrieval,in,Amsterdam)). Note that, in terms of structure and meaning, this differs from the previous one. Using brackets in the textual representation, these index expressions can be written as conference on (IT in (Belgium)) and conference on (IT) in (Belgium).

A parsing algorithm for index expressions is provided in [5]. It distinguishes connectors with high and low priority. The priority of a connector either leads to a broadening or a deepening of the structure. This is called structure detection. In tests performed on the titles of the CACM document collection it was found that the algorithm produced well-formed structures in approximately ninety percent of the cases. The erroneously parsed index expressions, however, might not cause severe problems since non-sensical descriptors are unlikely to be selected by users.

Index expressions can be broken down into subexpressions. Roughly speaking, the subexpressions of an index expression are all index expressions that it contains. In [15], several notions of subexpressions are provided. Direct subexpressions can be obtained from an index expression by removing a single leaf or, if possible, the head. All subexpressions can be obtained by repeating this defoliation several times.

Index expression add(add(conference,on,retrieval),in,Amsterdam). has two direct subexpressions. The first, add(conference,on,retrieval) is obtained by removing its rightmost subexpression. The second, add(conference,in,Amsterdam), is obtained by removing its leftmost subexpression. These, in turn, have a number of direct subexpressions, all consisting of a single term. These single terms, for instance conference, are no direct subexpressions of the original index expression. However, they are (general) subexpressions of the original index expression.

Subexpressions are used to build a fine-grained navigational network of index expressions. This network, called a lithoid, is based on an initial set of index expressions. In case of the stratified architecture, the union of characterisations constitutes this initial set. Uniting index expressions is done by term sharing. The lithoid consists of nodes and links. Each node is an index expression. The index expressions included in a lithoid are all index expressions from the characterisations of the initial set of documents plus all their subexpressions. The links of the lithoid connect index expressions. In order to obtain fine-grained links, nodes are connected with their direct subexpressions. This is illustrated in figure 2.

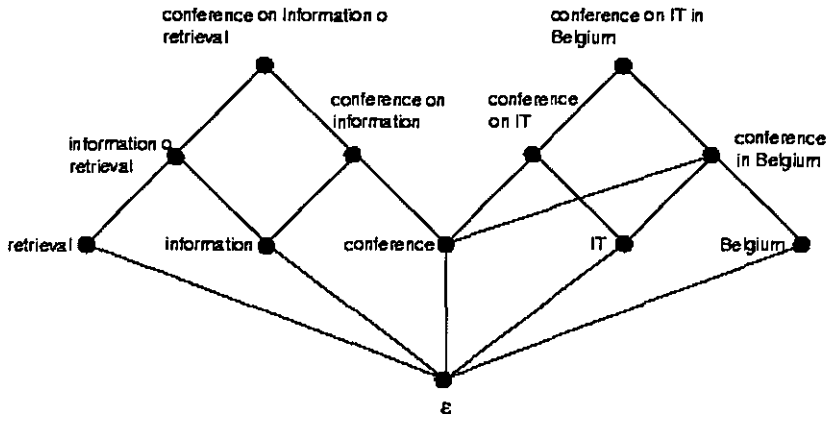


Figure 2:Example lithoid

Lithoids form networks consisting of nodes (index expressions) that are connected through links. Therefore, lithoids can be navigated in order to formulate queries.

3 Query by Navigation

Formulating an information need aims at finding a descriptor that properly describes it. Query by Navigation (QBN) ([6,4]) is a navigational way of query formulation in the stratified architecture for index expressions. By structurally navigating in the hyperindex, users formulate a query. During QBN, documents may be explored by transfers between hyperindex and base layer.

QBN identifies two types of actions: navigational actions in the hyperindex and beam operations for traveling from the hyperindex to the base layer and vice versa.

3.1 Navigating the Hyperindex

QBN starts by selecting a single node (index expression) in the hyperindex. The current node in the hyperindex is called the focus. The user may continue navigation by selecting one of the neighbours of the focus. The selected neighbour then becomes the focus and the selection process is repeated. Navigation thus essentially is a repetitive selection of neighbours. Navigation ends if a satisfactory index expression is reached. That is, if documents that satisfy the information need have been found.

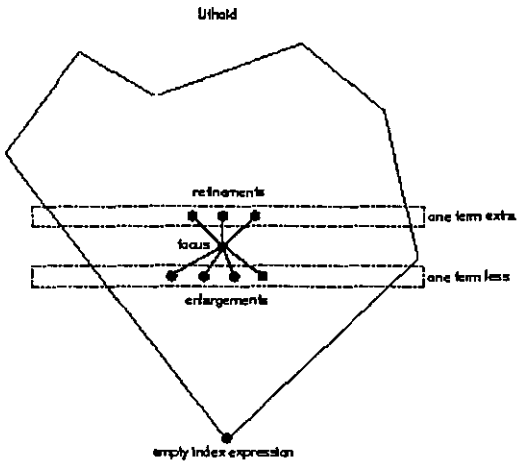


Figure 3: Neighbours in Query by Navigation

QBN exploits the special form of the hyperindex, i.e. a lithoid, by allowing fine-grained navigation steps. In figure 3, an abstract picture of a lithoid is given in which one of the nodes is marked as focus. The neighbours of the focus depict the direct choices for QBN. They thus give an overview of the concepts available for selection.

In a lithoid, the neigbours of a node are either refinements, direct superexpressions, or enlargements, direct subexpressions. Refinements, residing directly above the focus in the lithoid, denote more specific concepts. For example, conference on (IT) in (Belgium) is a refinement of both conference on IT and conference in Belgium. Since refinements contain one node more than the focus, they denote the smallest possible more specific concepts. This guarantees the fine-grained nature of the navigation steps in QBN. By selecting a refinement, the user formulates his need more concisely.

Enlargements denote less specific concepts than the focus. In fact, they denote a subconcept of the focus. For example, IT is an enlargement of conference on IT, which, in turn, is an enlargement of conference on (IT) in (Belgium). By selecting an enlargement, the user thus obtains a broader description. Enlargements can, for instance, be selected in order to recover from a previously selected refinement in order to change direction in the hyperindex. In this way, shifts in interests are fluently dealt with.

Navigation in the hyperindex consists of selecting a number of refinements and enlargements. In addition, the user can transfer himself from the hyperindex to the base layer to inspect actual documents. If required, the user can transfer himself back to the hyperindex, where navigation can be resumed.

3.2 Beaming between Layers

Inspecting actual documents is important for a user to ascertain the relevance of documents (searching) and for concept learning (exploration). Inspecting documents is enabled by an operation, beam down, that transfers the user from hyperindex to hyperbase (see figure 1).

By traveling the beam relation downward, the user is presented with the documents relevant to the focus in the hyperindex. The user can inspect these documents, and, if links between documents are available in the hyperbase, browse through them. If satisfied, the user can end the QBN session. In searching, for example, this may be the case if the user has satisfied his information need by rendering some relevant documents. In exploration, this may happen if the user estimates his knowledge of the field of interest is now sufficient.

In addition, the user may transfer himself back to the hyperindex (beam up) to resume navigation. However, since the characterisation of a document may contain several index expressions, a document can be linked to several nodes in the hyperindex. Therefore, the target node for the beam up may not be clearly defined. This problem, called ambiguity, may lead to user disorientation, especially when the user first follows a few browsing steps. Therefore, the INN system provides a rather basic back-button which guarantees that the user returns in the hyperindex at the same node where he left it.

4 INN System

The INN system forms an intermediary between users and the WWW. It makes use of existing search engines to access information on the WWW. The overall architecture of the INN system is sketched in figure 4.

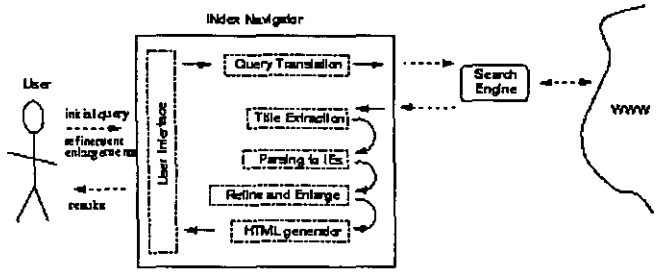


Figure 4: Overview of INN architecture

The path of control of the INN system is as follows. After the user has formulated an initial query, this is translated into the query language of the selected search engine. From the documents returned by the search engine, the titles are stripped. These titles are then parsed, so that index expressions are obtained. Refinements and enlargements in the parsed titles are computed and presented in a HTML page. If the user selects one of the navigational options, a refinement or an enlargement, the process is repeated with a new focus. The user may also go to the documents about one of the presented topics (beam down). Those documents then constitute the result of the system.

By providing an example session, we will discuss the workings of the INN system in more detail.

4.1 Getting Started

The initial screen of the INN system is given in figure 5. The components of this screen are explained below.

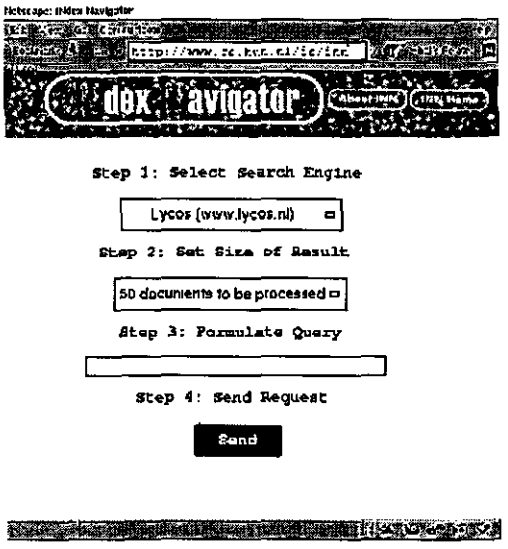


Figure 5: Initial Screen of INN

The About INN button provides information about the background and workings of the INN system. For example, the idea of QBN, using refinements and enlargements, is explained. The INN Home button leads the user to the home page of the INN system.

To start navigating the WWW, four steps have to be followed by the user.

1. Select Search Engine.
   First, a search engine has to be selected from a list. In figure 5, two well-known search engines, Alta Vista and Lycos, are available as well as two Dutch engines, Zoek and Ilse.

Other search engines can easily be incorporated. This also means that different information spaces can be navigated and explored via the INN system. From figure 4, one can see that the only changes needed are in the communication between the INN system and the search facility used to access the information space. The user query needs to be translated into a form that the search facility supports.

For textual information, this is mostly keyword-based. Since this is already supported, it means that the query translation part need not be changed. The only part that might need changes then is the title extractor. Since most search facilities clearly mark titles in their output, modifying the title extractor is a rather easy task.

2. Set Size of Result Set.
   Second, the size of the result set produced by the search facility must be set. That is, the number of documents that will be used for producing the overview should be specified. In this way, the user can steer the coverage of the overview. In addition, the user gains control over the response time of the system.

The size of the result set can be adapted for each step during navigation. This is a nice property since a larger query generally means that less documents are returned. By increasing the size of the result set, this effect can be overcome.

3. Provide Initial Query.
   Third, an initial query has to be provided. It is interpreted as index expression. The initial query may be of any size. However, it is recommended that a small initial query is provided in order to start with a broad overview.
4. Send Request.
   Finally, the user sends his request by clicking on the send button.

The query is fed to a search engine and the resulting documents are processed. In our example session, the user types the query retrieval.

4.2 Navigating on the Fly

The next page shown to the user, see figure 6, gives an overview of the navigational options. It consists of four parts, which are generated on the fly.

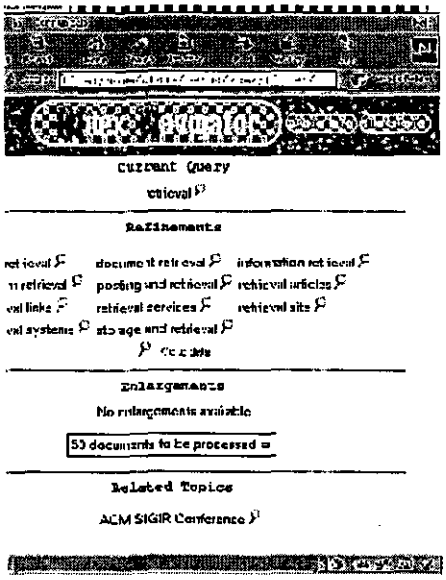


Figure 6: Overview Screen of INN

Focus.
The previously given query, i.e. retrieval, functions as current point in the hyperindex (focus). By clicking on the magnifying glass next to the focus, the user goes directly to the relevant documents (beam down). This option is offered for all descriptors.

Refinements.
Next, all refinements of the focus are given. For the example query, refinements include data retrieval, information retrieval, retrieval links, and storage and retrieval. The refinement give an overview of the topic retrieval.

Refinements are computed within the titles of the documents that were returned by the search facility. Every direct superexpression of the focus that is contained in any of the titles is listed.

Enlargements.
The enlargements of the focus are computed by defoliation ([15]). An enlargement of an index expression is obtained by removing a single leaf and its connector. In addition, if the root of the focus has only one subexpression, the root (and its connector) can also be removed to obtain an enlargement.

The empty index expression is not included in the INN system since it bears no information. This means that single terms have no enlargements in the INN system. Since the example query is a single term, figure 6 contains no enlargements.

Related topics.
Some of the titles do not (literally) contain the focus and thus do not lead to refinements. However, since they were rendered by the search facility, they may very well be relevant to the user. Therefore, the top 10 (according to the search engine's relevance estimates) documents are included. By clicking on the title, direct access to the document is provided. In addition, the magnifying glass is also given which uses the search facility to render documents that are related (about) the title. The selected search engine is used for this.

For the example query, a reference to a page about the Sigir conference on IR is given. Clearly, this topic is related.

4.3 Beaming between Layers

When a satisfactory descriptor has been reached or when an unknown concept is arrived at, the user can transfer himself to the documents that are about that descriptor (beam down). In this way, the user is enabled to see if the documents satisfy his information need and he can learn what the concept is about. The back button of the browser enables the user to beam up again.

In the presentation of the documents, the result of the search engine is included as a frame. In this frame, the user can use the facilities offered by the selected search engine. In addition, the user is enabled to return to the INN homepage to start a new navigation session.

4.4 Relation with CQE

In [12], the Condorcet Query Engine (CQE) is presented, a query engine for coordinated index terms. As the INN, CQE uses structured descriptors. Differences, however, are prominent. For instance, the coordinated concepts of the CQE reside in an ontology. The CQE is thus domain specific and requires careful maintenance of the used knowledge base.

In addition, no nested coordinated concepts are allowed in CQE. Therefore, we claim that step-wise refined descriptors are better supported by the INN. Since the CQE is only available as a prototype with a restricted example document space, a pragmatic comparison with the INN is not yet feasible. Whereas the CQE seems suitable for restricted domains and expert users, offering a well-motivated approach, the INN system provides a general dynamic interface for the WWW.

5. Conclusions

In this article, we showed how navigational tools for searching and exploring large and dynamic information spaces can be based on Query by Navigation. The size and dynamic nature of the information space, in our case the WWW, was coped with by creating the required part of the navigational network on the fly.

The contributions of this article are the following. First, we showed how our tool, the INdex Navigator, is grounded in the theory about navigational networks for index expressions. Second, we provided insight in the workings of the tool. As added new functionality in relation to similar tools, we included short cuts to related topics,

Further research can be directed in a number of ways. First, a graphical user interface may enhance the feeling of navigation. An good overview of graph visualisation techniques and applications can be found in [7].

Second, extracting index expressions from the contents of documents enhances the effectiveness of the tool as well. This may require additional linguistic knowledge in the form of, for instance, a part-of-speech tagger, extractor, and parser ([2]).

Third, the log file containing user actions that is created can be examined. This may lead to insight in the (relative) frequencies of the use of different actions (e.g. refinements and enlargements). In addition, it may be possible to obtain patterns in user behaviour, based on which user support during navigation may be supplied. User support may then aim at decreasing the user's cognitive load by filtering out topics that are likely to be irrelevant or at decreasing the expected search length by providing short cuts to expected destinations.

Bibliography

1 M. Agosti, G. Gradenigo, and P.G. Marchetti.
A Hypertext Environment for Interacting with Large Textual Databases.
Information Processing and Management, 28(3):371-387, 1992.

2 A.T. Arampatzis, T. Tsoris, C.H.A. Koster, and Th.P. van der Weide.
Phrase-based Information Retrieval.
Information Processing & Management, 34(6):693-707, December 1998.

3 N.J. Belkin, R.N. Oddy, and H.M. Brooks.
ASK for information retrieval. Part I. Background and theory.
In Journal of Documentation, volume 38, pages 61-71, 1982.

4 F.C. Berger.
Navigational Query Construction in a Hypertext Environment.
PhD thesis, Department of Computer Science, University of Nijmegen, September 1998.

5 P.D. Bruza.
Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia. PhD thesis, University of Nijmegen, Nijmegen, The Netherlands, 1993.

6 P.D. Bruza and Th.P. van der Weide.
Stratified Hypermedia Structures for Information Disclosure.
The Computer Journal, 35(3):208-220, 1992.

7 Herman, I. and Melancon, G. and Marshall, M.S.
Graph Visualisation in Information Visualisation.
In B. Falcidieno and J. Rossignac, editors, Proceedings of Eurographics '99, Aire-la-Ville, 1999.

8 D. Lucarella and Z. Zanzi.
Information Retrieval from Hypertext: An Approach using Plausible Inference.
Information Processing & Management, 29(3):299-312, 1993.

9 I. Ounis and M. Pasca.
RELIEF: Combining Expressiveness and Rapidity in one System.
In W.B. Croft, A. Moffat, C.J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 266-274, Melbourne, Australia, August 1998. ACM Press.

10R. Ragas.
The effect of linguistic form on user's behaviour and performance in free-text retrieval.
Master's thesis, Univerity of Nijmegen, Nijmegen, The Netherlands, 1996.

11D.R. Swanson.
Historical Note: Information Retrieval and the Future of an Illusion.
Journal of the American Society for Information Science, 32:92-98, 1988.

12Vet, P.van der and Mars, J.I.
CQE: a query engine for coordinated index terms.
Journal of the American Society for Information Systems, 50:485-492, 1999.

13J.A. Waterworth and M.H. Chignell.
A Model for Information Exploration.
Hypermedia, 3(1):35-58, 1991.

14R. Wilkinson and M. Fuller.
Integrated Information Access via Structure.
In M. Agosti and A. Smeaton, editors, Hypertext and Information Retrieval, pages 257 - 271, Boston, U.S.A, 1996. Kluwer.

15 B.C.M. Wondergem, P. van Bommel, and Th.P. van der Weide.
Nesting and Defoliation of Index Expressions for Information Retrieval.
Knowledge and Information Systems, 1999.

Supporting User Adaptation in Adaptive Hypermedia Applications

Hongjing Wu, Geert-Jan Houben, Paul De Bra

Department of Computing Science
Eindhoven University of Technology
PO Box 513, 5600 MB Eindhoven
the Netherlands

phone: +31 40 2472733
fax: +31 40 2463992
email: {hongjing,houben,debra}@win.tue.nl

Abstract

A hypermedia application offers its users a lot of freedom to navigate through a large hyperspace. The rich link structure of the hypermedia application can not only cause users to get lost in the hyperspace, but can also lead to comprehension problems because different users may be interested in different pieces of information or a different level of detail or difficulty. Adaptive hypermedia systems (or AHS for short) aim at overcoming these problems by providing adaptive navigation support and adaptive content. The adaptation is based on a user model *that represents relevant aspects about the user.*

At the Eindhoven University of Technology we developed an AHS, named AHA [DC98]. To describe its functionality and that of future adaptive systems we also developed a reference model for the architecture of adaptive hypermedia applications, named AHAM (for Adaptive Hypermedia Application Model) [DHW99]. In AHAM knowledge is represented through hierarchies of large composite abstract concepts as well as small atomic ones. AHAM also divides the different aspects of an AHS into a domain model *(DM), a* user model *(UM) and* an adaptation model *(AM). This division provides a clear separation of concerns when developing an adaptive hypermedia application.*

In this paper, we concentrate on the user modeling aspects of AHAM, but also describe how they relate to the domain model and the adaptation model. Also, we provide a separation between the adaptation rules an author or system designer writes (as part of the adaptation model) and the system's task of executing these rules in the right order. This distinction leads to a simplification of the author's or system designer's task to write adaptation rules. We illustrate authoring and adaptation in by some examples in the AHS AHA.

Keywords: adaptive hypermedia, user modeling, adaptive presentation, adaptive navigation, hypermedia reference model

1. Introduction

Hypermedia systems, and Web-based systems in particular, are becoming increasingly popular as tools for user-driven access to information. Hypermedia applications typically offer users a lot of freedom to navigate through a large hyperspace. Unfortunately, this rich link structure of the hypermedia application causes some serious usability problems:

- A typical hypermedia system presents the same links on a page, regardless the path a user followed to reach this page. When providing navigational help, e.g. through a map (or some fish-eye view) the system does not know which part of the link structure is most important for the user. The map cannot be simplified by filtering (or graying) out links that are less relevant for the user. Not having personalized maps is a typical *navigation problem* of hypermedia applications.
- Navigation in ways the author did not anticipate also causes *comprehension problems*: for every page the author makes an assumption about the foreknowledge the user has when accessing that page. However, there are too many ways to reach a page to make it possible for an author to anticipate all possible variations in foreknowledge when a user visits that page. A page is always presented in the same way. This often results in users visiting pages containing a lot of redundant information and pages that they cannot fully understand because they lack some expected foreknowledge.

Adaptive hypermedia systems (or AHS for short) aim at overcoming these problems by providing adaptive navigation support and adaptive content. Adaptive hypermedia is a recent area of research on the crossroad of hypermedia and the area of user-adaptive systems. The goal of this research is to improve the usability of hypermedia systems by making them personalized. The personalization or adaptation is based on a *user model* that represents relevant aspects about the user. The system gathers information about the user by observing the use of the application, and in particular by observing the *browsing* behavior of the user.

Many adaptive hypermedia systems exist to date. The majority of them are used in educational applications, but some are used for on-line information systems, on-line help systems, information retrieval systems, etc. An overview of systems, methods and techniques for adaptive hypermedia can be found in [B96]. At the Eindhoven University of Technology we developed an AHS system [DC98] out of Web-based courseware for an introductory course on hypermedia. In this system, called AHA, knowledge is represented with the same granularity as content: at the page level. In earlier versions of AHA, the user's knowledge about a given concept was a binary value: *known* or *not known*. The current version supports a more sophisticated representation in the sense that the knowledge level is represented by a *percentage*: reading a page can lead to an increase (or decrease) of the percentage. As part of the redesign process for AHA we have developed a reference model for the architecture of adaptive hypermedia applications: AHAM (for **A**daptive **H**ypermedia **A**pplication **M**odel) [DHW99], which is an extension of the Dexter hypermedia reference model [HS90, HS94]. AHAM acknowledges that doing "useful" and "usable" adaptation in a given application depends on three factors:

- The application must be based on a *domain model*, describing how the information content of the application (or "hyperdocument") is structured. This model must indicate what the relationship is between the high (and low) level concepts the application deals with, and it must indicate how concepts are tied to information fragments and pages.
- The system must construct and maintain a fine-grained *user model* that represents a user's preferences, knowledge, goals, navigation history and possibly other relevant aspects. The system can learn more about the user by observing the user's behavior. The user's knowledge is represented using the concepts from the domain model.
- The system must be able to adapt the presentation (of both content and link structure) to the reading and navigation style the user prefers and to the user's knowledge level. In order to do so the author must provide an *adaptation model* consisting of *adaptation rules*, for instance indicating how relations between concepts influence whether it will be desirable to guide the user towards or away from pages about certain concepts. Most AHS will offer a default adaptation model, relieving the author from explicitly writing these rules. In the original definition of AHAM [DHW99] we used the terms *teaching model* (TM) and *pedagogical rules*. These terms stem from the primary application of AHS's which is in education.

The key elements in AHAM are thus the *domain model* (DM), *user model* (UM) and *adaptation model* (AM). This division of adaptive hypermedia applications provides a clear separation of concerns when developing an adaptive hypermedia application.

The main shortcoming in many current AHS is that these three factors or components are not clearly separated:

- The relationship between pages and concepts is sometimes too vague (e.g. in [PDS98]). When an author decides that two pages each represent 30% of the same concept, there is no way of inferring whether together they represent 30%, 60% of the concept or any value in between. On the other hand systems like AHA [DC98] the relation between pages and concepts is strictly one-to-one, which leads to a very fragmented user model without high-level concepts.
- The adaptation rules can often not be defined at the conceptual level but only at the page level. In AHA [DC98], ELM-ART [BSW96a] and Interbook [BSW96b] for instance the destination of a link is (in almost all cases) a fixed page, described through a plain HTML anchor tag. (The "teach me" button in Interbook is an exception.)
- There may be a mismatch between the high level of detail in the user model and the low reliability of the information on which an AHS must update that user model. The basic information available to most AHS is the time at which a user requests a page (through a WWW-browser). Many educational AHS compensate for the unreliable event information by offering (multiple-choice) tests. A few systems, including AHA [DC98], capture reading time by logging both requests for pages and the time at which the user leaves a page (even when jumping to a different Web-site).

In this paper we focus on the user modeling aspects of AHAM and the use of adaptation rules to generate adaptive presentations and to update the user model. We extend the results of [WHD99b] by separating adaptation rules from the specification of the execution of these rules.

This paper is organized as follows. In Section 2 we describe the AHAM reference model for adaptive hypermedia applications. In Section 3 we elaborate on user modeling and on the use of adaptation rules in AHAM, that is how to construct the user model, update the user model by observing the user's behavior, and how to make content adaptation and link adaptation depending on the user model. In Section 4 we use AHAM to describe the user modeling and adaptation features of the AHA system, before we conclude in Section 5.

2. AHAM, a Dexter-based Reference Model

In hypermedia applications the emphasis is always on the information nodes and on the link structure connecting these nodes. The Dexter model captures this in what it calls the Storage Layer. It represents a *domain model* DM, i.e. the author's view on the application domain expressed in terms of concepts.

In adaptive hypermedia applications the central role of DM is shared with a *user model* UM. UM represents the relationship between the user and the domain model by keeping track of how much the user knows about each of the concepts in the application domain.

In order to perform adaptation based on DM and UM an author needs to specify how the user's knowledge influences the presentation of the information from DM. In AHAM this is done by means of a *teaching model* TM consisting of *pedagogical rules*. In this paper we use the terms *adaptation model* (AM) and *adaptation rules* to avoid the association with educational applications. An adaptive engine uses these rules to manipulate link anchors (from the Dexter model's *anchoring*) and to generate what the Dexter model calls the *presentation specifications*. Like the Dexter model, AHAM focuses on the Storage Layer, the anchoring and the presentation specifications. Figure 1 shows the structure of adaptive hypermedia applications in the AHAM model.

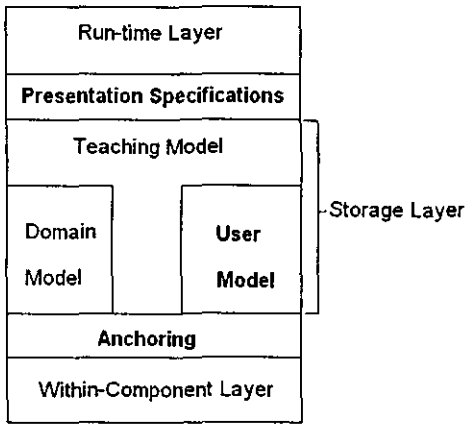


Figure 1: global structure of adaptive hypermedia applications.

In this section we present the elements of AHAM that we will use in Section 3 to illustrate the user modeling and adaptation.

2.1 The domain model

A *component* is an abstract notion in an AHS. It is a pair (uid, cinfo) where uid is a globally unique (object) identifier for the component and cinfo represents the component's information. A *component's information* consists of:

- A set of attribute-value pairs;
- A sequence of anchors (for attaching links);
- A presentation specification.

We distinguish two "kinds" of components: *concepts* and *concept relationships*. A *concept* is a component representing an abstract information item from the application domain. It can be either an *atomic concept* or a *composite concept*. An *atomic concept* corresponds to a fragment of information. It is primitive in the model (and can thus not be adapted). Its attribute and anchor values belong to the "Within-component layer" and are thus implementation dependent and not described in the model. A *composite component* has two "special" attributes:

- A sequence of children (concepts);
- A constructor function (to denote how the children belong together).

The children of a composite concept are all atomic concepts (then we call it a *page* or in typical hypertext terms a *node*) or all composite concepts. The composite concept component hierarchy must be a DAG (directed acyclic graph). Also, every atomic concept must be included in some composite concept. Figure 2 illustrates a part of a concept hierarchy.

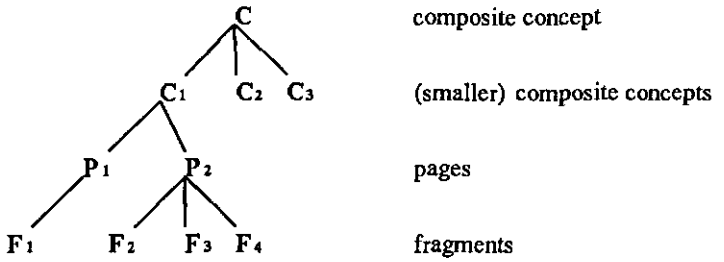


Figure 2: Example concept hierarchy.

An *anchor* is a pair (aid, avalue), where aid is a unique (object) identifier for the anchor within the scope of its component and avalue is an arbitrary value that specifies some location, region, item or substructure within a concept component.

Anchor values of atomic concepts belong to the (implementation dependent) Within-Component layer. Anchor values of composite concepts are identifiers of concepts that belong to that composite.

A *specifier* is a tuple (uid, aid, dir, pres), where uid is the identifier of a concept, aid is the identifier of an anchor, dir is a direction (FROM, TO, BIDIRECT, or NONE), and pres is a presentation specification.

A *concept relationship* is a component, with two additional attributes:

- A sequence of specifiers;
- A concept relationship type.

The most common type of concept relationship is the type **link**. This corresponds to the link components in the Dexter model, or links in most hypermedia systems. (Links typically have at least one FROM element and one TO or BIDIRECT element.) In AHAM we consider other types of relationships as well, which play a role in the adaptation. A common type of concept relationship is **prerequisite**. When a concept C_1 is a prerequisite for C_2 it means that the user should read C_1 before C_2. It does not mean that there must be a link from C_1 to C_2. It only means that the system somehow takes into account that reading about C_2 is not desired before some (enough) knowledge about C_1 has been acquired. Every prerequisite must have at least one FROM element and one TO element. Figure 3 shows a small set of (only binary) relationships, both prerequisites and links.

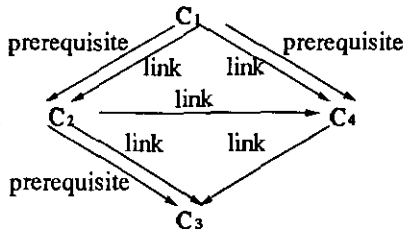


Figure 3: Example concept relationship structure.

The atomic concepts, composite concepts and concept relationships together form the *domain model* DM of an adaptive hypermedia application.

2.2 The user model

An AHS associates a number of *user model attributes* with each concept component of DM. For each user the AHS maintains a *table-like structure*, in which for each concept the attribute values for that concept are stored. Section 3 describes the user model in detail. For now it suffices to know that because of the relationships between *abstract* concepts and *concrete* content elements like fragments and pages a user model may contain other attributes than simply a *knowledge level*. For instance, the user model may also store information about what a user has actually read about a concept or whether a concept is considered relevant for the user.

Since the user model consists of "named entities" for which we store a number of attribute/value pairs, there is no reason to limit these "entities" to *concepts* about which the *knowledge level* is stored and updated. Concepts can be used (some might say abused) to represent other user features, such as preferences, goals, background and hyperspace experience. For the AHS (or the AHAM model) the actual meaning of concepts is irrelevant.

2.3 The adaptation (teaching) model

The adaptation of the information content of a hyperdocument and of the link structure is based on a set of *rules*. These rules form the connection between DM, UM and the presentation (specification) to be generated [WHD99a].

We partition the rules into four groups according to the adaptation "steps" to which they belong. These steps are IU, UU-Pre, GA, and UU-Post. An algorithm applies rules in each group. IU is to initialize the user model, under control of *Initialize-UM*; UU-Pre is to update UM before generating the page, under control of *Update-UM-pre*; GA is to generate adaptation, under control of *Adaptation*; UU-Post is to update UM after generating the page, under control of *Update-UM-post*. The four algorithms control how the rules in each group work together. By this we mean that an algorithm will trigger applicable rules (in some order) until no more rules can be applied or until the application of rules would no longer incur any change to UM.

A *generic adaptation rule* is a rule in which (bound) variables are used that represent concepts and concept relationships. A *specific adaptation rule* uses concrete concepts from DM instead of variables. Other than that both types of rules look the same. The syntax of the permissible rules depends on the AHS. In Section 3 we give examples of adaptation rules, using an arbitrarily chosen syntax. In Section 4 we give examples of adaptation rules as they are implemented in the AHA system [DC98]. Generic adaptation rules are often system-defined, meaning that an author need not specify them. Such a rule may for instance define how the knowledge level of an arbitrary concept C_1 influences the relevance of other concepts for which C_1 is a prerequisite. Author-defined rules always take precedence over (conflicting) system-defined rules. (Some AHS do not provide the possibility for authors to define their own generic adaptation rules.) Specific rules always take precedence over generic rules.

While specific rules are typically used to create exceptions to generic rules they can also be used to perform some ad-hoc adaptation based on concepts for which DM does not provide a relationship. Specific adaptation rules must always be defined by the author.

The *adaptation model* AM of an AHS is the set of (generic and specific) adaptation rules.

An AHS does not only have a domain model, user model en adaptation model, but also an *adaptive engine*, which is a software environment that performs the following functions:

- It offers generic page selectors and constructors. For each composite concept the constructor is used to determine which page to display when the user follows a link to that composite concept. For each page the constructor is used for building the adaptive presentation of that page.
- It optionally offers a (very simple programming) language for describing new page selectors and -constructors. Generic and specific adaptation rules (from UU-pre and GA) are used during page selection and construction.
- It performs adaptation by executing the page selectors and constructors. This means selecting a page, selecting fragments, sorting them, maybe presenting them in a specific way, etc. It also means performing adaptation to links by manipulating link anchors depending on the state of the link (like enabled, disabled, hidden, etc.).
- It updates the user model (instance) each time the user visits a page. It does so by triggering the necessary adaptation rules in UU-post. The engine will thus set the knowledge value for each atomic concept of displayed fragments of the page to a value that depends on a configurable amount (this could be 1 by default but possibly overridden by the author). It determines the influence on the knowledge value for page- and composite concepts. It also maintains other attribute values for each concept.

The adaptive engine thus provides the implementation dependent aspects while DM, UM and AM describe the information and adaptation at the conceptual, implementation independent level. An *adaptive hypermedia application* is a 4-tuple (DM, UM, AM, AE), where DM is a domain model. UM is a user model, AM is a adaptation model, and AE is an adaptive engine.

3. User Modeling and Adaptation in AHAM

According to AHAM the AHS maintains a fine-grained user model that represents the state of the user's features not only at the page level but also at the abstract conceptual level. It offers the ability to consider navigation history and other relevant user aspects as part of the user model UM. The maintenance of the relevant user aspects in UM is achieved by the application of the adaptation rules that are part of the adaptation model AM.

3.1 Representation of user features using (attribute/value) pairs

By definition adaptive hypermedia applications reflect some features of the user in the user model. This model is used to express various visible aspects of the system that depend on the user and that are visible to that user. Brusilovsky [B96] states which aspects of the user can be taken into account when providing adaptation. Generally, there are five user features that are used by existing AHS:

- knowledge
- user goals
- background
- hyperspace experience
- preferences

Almost every adaptive presentation technique relies on the user's knowledge as a source of adaptation. The system has to recognize the changes in the user's knowledge state and update its user model accordingly. Often the user's knowledge is represented by an overlay model. This overlay model is based on a conceptual structure of the subject domain. Sometimes a simpler stereotype user model is used to represent the user's knowledge: this means that the user is classified according to some stereotype. As many adaptation techniques require a rather fine-grained approach, stereotype models are often too simple to provide adequate personalization and adaptation. Overlay models on the other hand are generally hard to initialize. Acceptable results are often achieved by combining stereotype and overlay modeling: stereotype modeling is used in the beginning to classify a new user and to set initial values for the overlay model; later a more fine-grained overlay model is used. Using the AHAM definition for user model, it is fairly straightforward how a user's knowledge state can be represented by associating a *knowledge value* attribute to each concept.

Apart from the concept's identifier (which may be just a name) a typical AHS will store not only a *knowledge value* for each concept, but also a *read* value which indicates whether (and how much) information about the concept has been read by the user, and possibly some other attribute values as well. While the model uses a table representation, implementations of AHS may use different data structures. For instance, a logfile can be used for the *read* attribute.

Table 1 illustrates the (conceptual) structure of a user model for a course on hypermedia: the concepts Xanadu and KMS were at least partially learnt. The concept WWW, consisting of two sub-parts, is partially learnt because WWW-page1 has been read but WWW-page2 has not been read. One can see that WWW must be a composite concept that is not a page, because it is already partially learnt while it has not been read at all.

concept name (uid)	Knowledge value	read	...
Xanadu	well learned	true	...
KMS	learned	true	...
WWW-page1	well learned	true	...
WWW-page2	not known	false	...
WWW	learned	false	...
...	...	...	...

Table 1: Example user model (instance).

The second kind of user feature is the user's goal. The user's goal or task is a feature that is related with the context of the user's working activities rather than with the user as an individual. The user's goal is the most volatile of all user features. It can be considered as a very important user feature for AHS. One representation of possible user goals uses a hierarchy (a tree) of tasks. Another representation of the user's current goal uses a set of pairs (Goal,Value), where Value is the probability that Goal is the current goal of the user. The latter representation perfectly matches the way in which AHAM models the user's state.

Two features of the user that are similar to the user's knowledge of the subject but that functionally differ from it, are the user's background and the user's experience in the given hyperspace. By background we mean all the information related to the user's previous experience *outside* the subject of the hypermedia system. By user's experience in the given hyperspace we mean how familiar is the user with the structure of the hyperspace and how easy can the user navigate in it. Again, these features can be modeled in AHAM using concepts' attribute/value pairs.

For different possible reasons the user can prefer some nodes and links over others or some parts of a page over others. This is used most heavily in information retrieval hypermedia applications. In fact in most adaptive information retrieval hypermedia applications preferences are the only information that is stored about the user. User preferences differ from other user model components, since in most cases they cannot be deduced by the system. The user has to inform the system directly or indirectly about the preferences. AHAM's attribute/value pairs can again be used to model the user's preferences.

From the above descriptions we can conclude that although a user model needs to represent (five) very different aspects of a user, all of these kinds of aspects can be implemented as sets of *concepts with associated attribute/value pairs*. For presentation purposes it is not necessary to treat these different kinds of aspects in a different way, but for implementation purposes it is often needed to treat these in different ways in adaptive hypermedia applications.

The knowledge value of a concept can be a Boolean, discrete or continuous value depending on the choice of the author (or the properties of the AHS). By using a Boolean value, the knowledge about the concept can be either *known* or *unknown*.

By using a discrete value the knowledge about the concept can be one of a small set of values, like *unknown*, *learnt*, *well learnt* or *well known*. By using continuous values from the range of [0..1], the value can more precisely describe the user's knowledge, and even describe the loss or decay of knowledge over time. In conclusion, AHAM's user model UM has enough expressive power to model all user features that current AHS take into account.

3.2 Changes in user features

In the previous subsection we discussed features that describe the user's state in the browsing process. Usually in adaptive hypermedia applications (as opposed to adaptable hypermedia applications, see [DHW99]), only the browsing behavior is observed in order to influence the adaptation. Basically, there are five ways in which the user features can change in an adaptive hypermedia application:

1. the user clicks on an anchor (and follows a link);
2. the user performs a test (explicitly);
3. information (about the user) is imported from an external testing system;
4. a user preference is (explicitly) set or declared by the user (initially);
5. a user preference is (automatically) inferred from the user's behavior.

Besides observing the browser behavior, the application can change the user features based on information that is explicitly imported from its environment or that is explicitly declared or implicitly inferred about the user's preferences.

These five different kinds of changes lead to five kinds of "rules" how to maintain the user features. The system can be made more *author centered* by including rules of types 2 and 3 (besides rules of type 1), while the application can become more *user centered* by including rules of types 4 and 5. It is also possible to choose a combination that suits the application.

3.3 Adaptation based on the user model

By maintaining the user model the system can infer how relevant aspects of the user change while the user is using the application and thus is using the adaptation. The adaptive engine realizes adaptive presentation and adaptive navigation (or link adaptation) according to the (adaptation) rules that are system-defined or written by the author and that depend on the user model.

Below we give a number of examples to show how adaptation rules are used to do adaptation. The syntax used for the rules is arbitrary and only exemplary. AHAM does not prescribe any specific syntax. Normally every AHS will provide its own syntax for defining adaptation rules.

Example 1 *For atomic concepts (fragments) let us assume that the presentation specification is a two-valued (almost Boolean) field, which is either "show" or "hide". When a page is being accessed, the following rule sets the visibility for fragments that belong to a "page" concept, depending on their "relevance" attribute-value.*

< access(C) and F IN C.children and F.relevance = true => F.pres := show >

Here we simplified things, by assuming that we can treat C.children as if it were a set, whereas it really is a sequence. It is common to execute rules for generating presentation specifications before generate the page, so it is in GA.

Example 2 *The following rules set the presentation specification for a specifier that denotes a link (source) anchor depending on whether the destination of the link is considered relevant and whether the destination has been read before. For simplicity we consider a link with just one source and one destination.*

< CR.type = link and CR.cinfo.dir[1] = FROM and CR.cinfo.dir[2] = TO and CR.ss[2].uid.relevant = true and CR.ss[2].uid.read = false => CR.ss[1].pres = GOOD >

< CR.type = link and CR.cinfo.dir[1] = FROM and CR.cinfo.dir[2] = TO and CR.ss[2].uid.relevant = true and CR.ss[2].uid.read = true => CR.ss[1].pres = NEUTRAL >

< CR.type = link and CR.cinfo.dir[1] = FROM and CR.cinfo.dir[2] = TO and CR.ss[2].uid.relevant = false => CR.ss[1].pres = BAD>

These rules say that links to previously unread but "relevant" pages are "GOOD". Links to previously read and "relevant" pages are "NEUTRAL" and links to pages that are not "relevant" are "BAD". In the AHA system [DC98] this results in the link anchors being colored blue, purple or black respectively. In ELM-ART [BSW96a] and Interbook [BSW96b] the links would be annotated with a green, yellow or red ball. We can consider the actual presentation (the coloring of the anchors) as belonging to the Run-time Layer and thus outside the scope of AHAM. However, should we opt to include the color preferences for GOOD, NEUTRAL and BAD links in the user model then the translation of the presentation specification to the color could still be described using a adaptation rule. These rules are in GA also.

3.4 Maintenance of user model

To record the reading history of the user and the evolution of the user's knowledge, the system updates the user model based on the observation of the user's browsing process. The rules that the author has defined in AM describe how to keep track of the evolution of the user's knowledge. For the application of adaptation rules we assume that the *FollowLink* operation from the Dexter (and thus AHAM) model's Run-time Layer results in a call to a *resolver function* for a given specifier. In AHAM the resolver translates the given specifier to the uid of a composite concept component that corresponds to a page, or to a set of such uid's. Which page exactly is selected depends on DM and UM. For the selected page an *accessor function* is called, according to the Dexter model, which returns the (page) concept component that corresponds to the resolved uid. Then the rules for presentation are executed, as shown in Subsection 3.3.

Example 3 *The following rule expresses that when a page is accessed the "read" user-model attribute for the corresponding concept is set to true:*

< access(C) => C.read := true>

This rule is in *UU-post*. It is the *Update-UM-post* that will trigger other rules that have *read* on their left-hand side in the same group.

Example 4 *The following rule expresses that when a page is "relevant" and it is accessed, the knowledge value of the corresponding concept becomes "well-learnt". This is somewhat like the behavior of Interbook [BSW96b].*

< access(C) and C.relevant = true => C.knowledge := well-learnt>

In Interbook, as well as in AHA [DC98], knowledge is actually updated before the page is generated. These rules thus are in *UU-pre*. At the end of Section 4 we shall describe why this option is chosen, and which problems it creates. In general one wishes to have the option to base some adaptation on the knowledge state *before* accessing a page and some adaptation on the knowledge state *after* reading the page.

Example 5 *The following rule expresses that after a user has taken a test about a concept C, his knowledge about concept C is changed (a rule of "type 2" from Subsection 3.2). Here, an action "test" is used that represents that a test has been taken. It is in UU-pre*

< test(C) and C.test > 60 => C.knowledge := known>

4. User Modeling and Adaptation in the AHA system

AHA [DC98] is a simple adaptive hypermedia system. We describe the properties of the version that is currently being used for two on-line courses and one on-line information kiosk, plus some features of the next version that is currently being developed.

- In AHA the *domain model* consists of three types of concepts: *abstract concepts*, *fragments* and *pages*. Concepts are loosely associated with (HTML) pages, not with fragments.
- The *user model* consists of:
  - Color preferences for link anchors which the user can customize. (These preferences result in "non-relevant" link anchors to be hidden if their color is set to black, or visibly "annotated" if their color is set to a non-black color, different from that of "relevant" link anchors.)
  - For each abstract concept, a *knowledge* attribute with percentage values. (100 means the concept is fully known). For pages and fragments there is no *knowledge* attribute value.
  - For each page, a Boolean *read* attribute. (*True* means the page was read, *false* means it was not read.) AHA actually logs access and reading times, but they cannot be used in a more sophisticated way in the current version. For abstract concepts and fragments there is no *read* attribute value.
- AHA comes with an *adaptation model* containing system-defined generic adaptation rules. It offers a simple language for creating author-defined specific adaptation rules (but no author-defined generic rules).

The *domain model* can only contain concept relationships of the types that are shown below. An author cannot define new types. The influence of these relationships on the adaptation and the user model updates is defined by system-defined generic adaptation rules. In AHA all rules are executed before generate the page and are triggered directly by a page access, thus eliminating the need for propagation.

- When a page is accessed, its *read* attribute in the user model is updated as follows (it is in *UU-pre*):

  < access(P) => P.read := true >

- The relationship type *generates* links a page to an abstract concept. A *generates* relationship between P and C means that reading page P generates knowledge about C (it is in *UU-pre*):

  < access(P) => C.knowledge := 100>

  This "generation" of knowledge in AHA is controlled by a structured comment in an HTML page:

  <!-- generates readme -->

  This example *generates* comment denotes that the concept readme becomes known when the page is accessed.

- The relationship type *requires* links a concept to a virtual composite concept that is defined by a (constructor which is a) Boolean expression of concepts. Although in principle this composite concept is unnamed, we shall use a "predicate" or "pseudo attribute of the page" to refer to it. *P.requires* is used as a Boolean attribute of which the value is always that of the corresponding Boolean expression. It is not a user model attribute as its value is always computed on the fly and not stored in the user model. A *requires* relationship is implemented using a structured comment at the top of an HTML page, e.g.:

  <!-- requires (readme and intro) -->

  This example expresses that this page is only considered relevant when the concepts readme and intro are both known (100%). In AHA, links to a page for which *requires* is *false* are considered BAD, and reading such a page generates less knowledge than reading a GOOD page. Below we give the rules in *GA* that determines how the link anchors will be presented. They are very similar to the rules in Example 2 (Subsection 3.3):

  < CR.type = link and CR.cinfo.dir[1] = FROM and CR.cinfo.dir[2] = TO and CR.ss[2].uid.requires = true and CR.ss[2].uid.read = false => CR.ss[1].pres = GOOD >

  < CR.type = link and CR.cinfo.dir[1] = FROM and CR.cinfo.dir[2] = TO and CR.ss[2].uid.requires =

true and CR.ss[2].uid.read = true => CR.ss[1].pres = NEUTRAL >

< CR.type = link and CR.cinfo.dir[1] = FROM and CR.cinfo.dir[2] = TO and CR.ss[2].uid.requires = false => CR.ss[1].pres = BAD>

- The relationship type *link* only applies to pairs of pages in AHA. "Page selectors" that exist in AHAM in general are thus not needed (or possible) in AHA.

AHAM allows author-defined specific adaptation rules only for the conditional inclusion of fragments in HTML pages. Structured HTML comments are used for specifying these rules. With a fragment F we can associate a "pseudo attribute" *requires* to indicate the condition, just like for whole pages. The syntax is illustrated by the following example:

```
<!-- if ( readme and not intro ) -->
... here comes the content of the fragment ...
<!-- else -->
... here is an alternative fragment ...
<!-- endif -->
```

AHA only includes fragments when their *requires* "attribute" is *true*.

The above examples illustrate that representing the actual functionality of an existing AHS in the AHAM model is fairly straightforward. The main reasons for using such a representation are to be able to compare different AHS, to possibly translate an adaptive hypermedia application from one AHS to another, and to identify potential problems or shortcomings in existing AHS.

We conclude this Section with an illustration of one specific shortcoming that we have found in both AHA [DC98] and Interbook [BSW96b]: the "new" knowledge values are calculated before generating the page (and in fact these systems do not support calculating knowledge values after generating a page at all). When a user requests a page, the knowledge generated by reading this page is already taken into account during the generation of the page. This has desirable as well as undesirable side-effects:

- When links to other pages become relevant *after* reading the current page it makes sense to already annotate the link anchors as relevant when presenting the page. Once a page is generated its presentation remains static while the user is reading it (and rightfully so). The new knowledge thus needs to be taken into account before the page is actually read.

- Pages contain information that becomes relevant or non-relevant depending on the user's knowledge. In some cases the relevance of a fragment may depend on the user having read the page that contains this fragment. This means that a fragment may be relevant the first time a page is visited and non-relevant thereafter, or just the other way round.

  By already taking into account the knowledge before the page is generated for the first time a different "first time version" becomes impossible to create. (Some readers may argue that having content that changes in this way may not be desirable in any case, but not having this possibility limits the general applicability of the AHS.)

5. Conclusions and Future Work

Over the past few years we have developed an AHS, mainly for use in courseware. We have come across a number of other AHS, with different interesting properties. As part of the redesign of AHA [DC98] we developed a reference model for AHS, named AHAM. The description of adaptive hypermedia applications in terms of this model has provided us with valuable redesign issues. The three most important ones are:

- The division of an adaptive hypermedia application into a *domain model*, *user model*, and *adaptation model* provides a clear separation of concerns and will lead to a better separation of orthogonal parts of the AHS functionality in the implementation of the next version of AHA. We believe that a system which supports this separation of concerns will not only result in a cleaner implementation, but also in a more usable authoring environment [WHD99a].
- In this paper we have described the *adaptation rules* in such a way that the rule definition is independent of the rule execution. This makes authoring easier.

- By representing AHA in the AHAM model we have identified another shortcoming: the lack of a two-phase application of rules. We found that this shortcoming is present in other AHS as well.

We deliberately based the AHAM model on the Dexter hypermedia reference model [HS90, HS94], to show that AHS are "true" hypermedia systems. In this paper we have concentrated on user modeling and adaptation. The description of these aspects at an abstract level sets AHAM apart from other descriptions of AHS that are too closely related to the actual implementation of these AHS.

In the near future we will develop a new version of the AHA system, in which the separation of *domain model*, *user model* and *adaptation model* will be more complete. We also plan an extended paper with a complete formal definition of AHAM, including a formal specification of a language for specifying adaptation rules.

References

[B96] Brusilovsky, P., "Methods and Techniques of Adaptive Hypermedia". User Modeling and User-Adapted Interaction, 6, pp. 87-129, 1996. (Reprinted in Adaptive Hypertext and Hypermedia, Kluwer Academic Publishers, pp. 1-43, 1998.)

[DC98] De Bra, P., Calvi, L., "AHA: a Generic Adaptive Hypermedia System". Proceedings of the Second Workshop on Adaptive Hypertext and Hypermedia, Pittsburgh, pp. 5-11, 1998.

[DHW99] De Bra, P., Houben, G.J., Wu, H., "AHAM: A Dexter-based Reference Model for Adaptive Hypermedia". Proceedings of ACM Hypertext'99, Darmstadt, pp. 147-156, 1999.

[HS90] Halasz, F., Schwartz, M., "The Dexter Reference Model". Proceedings of the NIST Hypertext Standardization Workshop, pp. 95-133, 1990.

[HS94] Halasz, F., Schwartz, M., "The Dexter Hypertext Reference Model". Communications of the ACM, Vol. 37, nr. 2, pp. 30-39, 1994.

[PDS99] Pilar da Silva, D., "Concepts and documents for adaptive educational hypermedia: a model and a prototype", Proceedings of the Second Workshop on Adaptive Hypertext and Hypermedia, Pittsburgh, pp. 33-40, 1998.

[WHD99a] Wu, H., Houben, G.J., De Bra, P., "Authoring Support for Adaptive Hypermedia", Proceedings ED-MEDIA'99, Seattle, pp. 364-369, 1999.

[WHD99b] Wu, H., Houben, G.J., De Bra, P., "User Modeling in Adaptive Hypermedia Applications", Proceedings InfWet99, Amsterdam, 1999.

Computing Science Reports **Department of Mathematics and Computing Science**
Eindhoven University of Technology

If you want to receive reports, send an email to: m.m.j.l.philips@tue.nl (we cannot guarantee the availability of the requested reports)

In this series appeared:

97/02	J. Hooman and O. v. Roosmalen	A Programming-Language Extension for Distributed Real-Time Systems, p. 50.
97/03	J. Blanco and A. v. Deursen	Basic Conditional Process Algebra, p. 20.
97/04	J.C.M. Baeten and J.A. Bergstra	Discrete Time Process Algebra: Absolute Time, Relative Time and Parametric Time, p. 26.
97/05	J.C.M. Baeten and J.J. Vereijken	Discrete-Time Process Algebra with Empty Process, p. 51.
97/06	M. Franssen	Tools for the Construction of Correct Programs: an Overview, p. 33.
97/07	J.C.M. Baeten and J.A. Bergstra	Bounded Stacks, Bags and Queues, p. 15.
97/08	P. Hoogendijk and R.C. Backhouse	When do datatypes commute? p. 35.
97/09	Proceedings of the Second International Workshop on Communication Modeling, Veldhoven, The Netherlands, 9-10 June, 1997.	Communication Modeling- The Language/Action Perspective, p. 147.
97/10	P.C.N. v. Gorp, E.J. Luit, D.K. Hammer E.H.L. Aarts	Distributed real-time systems: a survey of applications and a general design model, p. 31.
97/11	A. Engels, S. Mauw and M.A. Reniers	A Hierarchy of Communication Models for Message Sequence Charts, p. 30.
97/12	D. Hauschildt, E. Verbeek and W. van der Aalst	WOFLAN: A Petri-net-based Workflow Analyzer, p. 30.
97/13	W.M.P. van der Aalst	Exploring the Process Dimension of Workflow Management, p. 56.
97/14	J.F. Groote, F. Monin and J. Springintveld	A computer checked algebraic verification of a distributed summation algorithm, p. 28
97/15	M. Franssen	λP-: A Pure Type System for First Order Loginc with Automated Theorem Proving, p.35.
97/16	W.M.P. van der Aalst	On the verification of Inter-organizational workflows, p. 23
97/17	M. Vaccari and R.C. Backhouse	Calculating a Round-Robin Scheduler, p. 23.
97/18	Werkgemeenschap Informatiewetenschap redactie: P.M.E. De Bra	Informatiewetenschap 1997 Wetenschappelijke bijdragen aan de Vijfde Interdisciplinaire Conferentie Informatiewetenschap, p. 60.
98/01	W. Van der Aalst	Formalization and Verification of Event-driven Process Chains, p. 26.
98/02	M. Voorhoeve	State / Event Net Equivalence, p. 25
98/03	J.C.M. Baeten and J.A. Bergstra	Deadlock Behaviour in Split and ST Bisimulation Semantics, p. 15.
98/04	R.C. Backhouse	Pair Algebras and Galois Connections, p. 14
98/05	D. Dams	Flat Fragments of CTL and CTL*: Separating the Expressive and Distinguishing Powers. P. 22.
98/06	G. v.d. Bergen, A. Kaldewaij V.J. Dielissen	Maintenance of the Union of Intervals on a Line Revisited, p. 10.
98/07	Proceedings of the workhop on Workflow Management: Net-based Concepts, Models, Techniques and Tools (WFM'98) June 22, 1998 Lisbon, Portugal	edited by W. v.d. Aalst, p. 209
98/08	Informal proceedings of the Workshop on User Interfaces for Theorem Provers. Eindhoven University of Technology ,13-15 July 1998	

		edited by R.C. Backhouse, p. 180
98/09	K.M. van Hee and H.A. Reijers	An analytical method for assessing business processes, p. 29.
98/10	T. Basten and J. Hooman	Process Algebra in PVS
98/11	J. Zwanenburg	The Proof-assistemt Yarrow, p. 15
98/12	Ninth ACM Conference on Hypertext and Hypermedia Hypertext '98 Pittsburgh, USA, June 20-24, 1998 Proceedings of the second workshop on Adaptive Hypertext and Hypermedia.	Edited by P. Brusilovsky and P. De Bra, p. 95.
98/13	J.F. Groote, F. Monin and J. v.d. Pol	Checking verifications of protocols and distributed systems by computer. Extended version of a tutorial at CONCUR'98, p. 27.
98/14	T. Verhoeff (artikel volgt)	
99/01	V. Bos and J.J.T. Kleijn	Structured Operational Semantics of χ , p. 27
99/02	H.M.W. Verbeek, T. Basten and W.M.P. van der Aalst	Diagnosing Workflow Processes using Woflan, p. 44
99/03	R.C. Backhouse and P. Hoogendijk	Final Dialgebras: From Categories to Allegories, p. 26
99/04	S. Andova	Process Algebra with Interleaving Probabilistic Parallel Composition, p. 81
99/05	M. Franssen, R.C. Veltkamp and W. Wesselink	Efficient Evaluation of Triangular B-splines, p. 13
99/06	T. Basten and W. v.d. Aalst	Inheritance of Workflows: An Approach to tackling problems related to change, p. 66
99/07	P. Brusilovsky and P. De Bra	Second Workshop on Adaptive Systems and User Modeling on the World Wide Web, p. 119.
99/08	D. Bosnacki, S. Mauw, and T. Willemse	Proceedings of the first international syposium on Visual Formal Methods - VFM'99
99/09	J. v.d. Pol, J. Hooman and E. de Jong	Requirements Specification and Analysis of Command and Control Systems
99/10	T.A.C. Willemse	The Analysis of a Conveyor Belt System, a case study in Hybrid Systems and timed μCRL, p. 44.
99/11	J.C.M. Baeten and C.A. Middelburg	Process Algebra with Timing: Real Time and Discrete Time, p. 50.
99/12	S. Andova	Process Algebra with Probabilistic Choice, p. 38.
99/13	K.M. van Hee, R.A. van der Toorn, J. van der Woude and P.A.C. Verkoulen	A Framework for Component Based Software Architectures, p. 19
99/14	A. Engels and S. Mauw	Why men (and octopuses) cannot juggle a four ball cascade, p. 10
99/15	J.F. Groote, W.H. Hesselink, S. Mauw, R. Vermeulen	An algorithm for the asynchronous *Write-All* problem based on process collision*, p. 11.
99/16	G.J. Houben, P. Lemmens	A Software Architecture for Generating Hypermedia Applications for Ad-Hoc Database Output, p. 13.
99/17	T. Basten, W.M.P. v.d. Aalst	Inheritance of Behavior, p.83
99/18	J.C.M. Baeten and T. Basten	Partial-Order Process Algebra (and its Relation to Petri Nets), p. 79
99/19	J.C.M. Baeten and C.A. Middelburg	Real Time Process Algebra with Time-dependent Conditions, p.33.
99/20	Proceedings Conferentie Informatiewetenschap 1999 Centrum voor Wiskunde en Informatica 12 november 1999, p.98	edited by P. de Bra and L. Hardman
00/01	J.C.M. Baeten and J.A. Bergstra	Mode Transfer in process Algebra, p. 14
00/02	J.C.M. Baeten	Process Algebra with Explicit Termination, p. 17.
00/03	S. Mauw and M.A. Reniers	A process algebra for interworkings, p. 63.
00/04	R. Bloo, J. Hooman and E. de Jong	Semantical Aspects of an Architecture for Distributed Embedded Systems*, p. 47.

00/05	J.F. Groote and M.A. Reniers	Algebraic Process Verification, p. 65.
00/06	J.F. Groote and J. v. Wamel	The Parallel Composition of Uniform Processes wit Data, p. 19
00/07	C.A. Middelburg	Variable Binding Operators in Transition System Specifications, p. 27.
00/08	I.D. van den Ende	Grammars Compared: A study on determining a suitable grammar for parsing and generating natural language sentences in order to facilitate the translation of natural language and MSC use cases, p. 33.
00/09	R.R. Hoogerwoord	A Formal Development of Distributed Summation, p. 35
00/10	T. Willemse, J. Tretmans and A. Klomp	A Case Study in Formal Methods: Specification and Validation on the OM/RR Protocol, p. 14.
00/11	T. Basten and D. Bošnački	Enhancing Partial-Order Reduction via Process Clustering, p. 14
00/12	S. Mauw, M.A. Reniers and T.A.C. Willemse	Message Sequence Charts in the Software Engineering Process, p. 26
00/13	J.C.M. Baeten, M.A. Reniers	Termination in Timed Process Algebra, p. 36
00/14	M. Voorhoeve, S. Mauw	Impossible Futures and Determinism, p. 14
00/15	M. Oostdijk	An Interactive Viewer for Mathematical Content based on Type Theory, p. 24.
00/16	F. Kamareddine, R. Bloo, R. Nederpelt	Characterizing λ-terms with equal reduction behavior, p. 12
00/17	T. Borghuis, R. Nederpelt	Belief Revision with Explicit Justifications: an Exploration in Type Theory, p. 30.
00/18	T. Laan, R. Bloo, F. Kamareddine, R. Nederpelt	Parameters in Pure Type Systems, p. 41.
00/19	J. Baeten, H. van Beek, S. Mauw	Specifying Internet applications with *DiCons*, p. 9