

Van papier tot morgen
over de problemen van duurzame digitale informatie

HT de Beer
0491874
H.T.d.Beer@student.tue.nl

Eindhoven, 15 december 2003

Inhoudsopgave

1	Inleiding	2
2	Duurzame digitale informatie	4
3	Het probleem van duurzame digitale informatie	6
3.1	Een praktisch probleem: hard- en software	6
3.2	Is digitale informatie wel te vertrouwen?	7
3.3	Wie wat bewaart, heeft wat	8
3.4	Informatie is meer dan enkel inhoud	9
4	Strategieën ter vergroting der duurzaamheid	11
4.1	Afgedrukt en wel, maar niet digitaal	11
4.2	Oudheidkundige informatica: computermusea	12
4.3	Na het een komt het ander: migratie	13
4.4	Waren standaarden maar standaard	14
4.5	Informatie over informatie	15
4.6	Emulatie	16
4.7	En de beste strategie is ... Een aanbeveling	16
5	Conclusie	18
	Literatuur	20

Hoofdstuk 1

Inleiding

We schrijven het jaar negentienhonderdneenennegentig, met het naderen van het volgende jaar stijgt de spanning. Het jaar tweeduizend was aanstaande en het mooie ronde getal inspireerde, net als bij vergelijkbare jaartallen in de geschiedenis, mensen van allerlei slag tot het doen van voorspellingen, vaak doemscenario's, maar ook Utopia werd aanstaande geroepen. Opmerkelijk echter was het grote aandeel van technici en beleidsmakers in de negatieve voorspellingen, normaal toch nuchtere lieden.

In de laatste decennia van de vorige eeuw was langzaam maar zeker de hele maatschappij doordrongen geraakt van de ICT. De technische mogelijkheden waren ten opzichte van het begin van de informatisering zo sterk gegroeid, dat de problemen uit het verleden niet meer leken te bestaan. Die problemen waren immers al lang geleden opgelost. Zo ook het probleem van het kleine geheugen, besparen op geheugengebruik was het devies destijds, door bijvoorbeeld een datum met maar twee getallen te representeren, dat kon toch geen kwaad?

Het welbekende 'jaar-tweeduizend'-probleem liet de maatschappij de adem inhouden op nieuwjaarsdag tweeduizend, gelukkig bleef het bij spanning alleen, maar een belangrijke les kan hieruit geleerd: *oplossingen van vandaag zijn problemen van morgen*. Nu gaat deze les natuurlijk niet op in het generaal, doch het is van belang om problemen op te lossen zonder daarmee een nieuw probleem te creëren. Oftewel het is belangrijk om duurzame oplossingen te bedenken.

Nu lijken informatica en duurzaamheid weinig met elkaar te maken te hebben, maar ook in de informatica moeten oplossingen en beslissingen genomen die veel langer hun invloed doen gelden dan een enkele generatie hard- en software. De korte representatie van de datum is daar een voorbeeld van.

Als er iets is wat de eeuwen doorstaat en ook moet doorstaan, dan is dat wel informatie: boeken, kunst, cultuur, bouwwerken, kranten, enz., oftewel ons culturele erfgoed. Al deze informatie wordt bewaard, er zijn zelfs speciale in-

stellingen voor in het leven geroepen zoals musea, bibliotheken en archieven. Dat het behouden van ons erfgoed belangrijk is, dient geen uitleg, toch is er de laatste jaren een belangrijk verschil ontstaan met vroeger tijden: informatie is tegenwoordig *digitaal*.

Na jarenlange digitalisering van informatie is het besef gerezen dat het bewaren van digitale informatie een probleem is. Ook hier zijn het beslissingen en oplossingen voor problemen uit het verleden, waarvoor de informatica zich gesteld zag, die nu zelf het probleem vormen. Dat probleem is het onderwerp van dit document. De vraag die centraal staat is *hoe het gesteld staat met de duurzaamheid van informatiesystemen*. Het antwoord wordt gezocht door eerst dieper in te gaan op wat duurzame digitale informatie en duurzame informatiesystemen nu precies zijn, dat is het onderwerp van het volgende hoofdstuk. Daarna, in hoofdstuk drie, worden de problemen van duurzame digitale informatie besproken. Waarom is het zo moeilijk om digitale informatie langer te bewaren dan enkele jaren? Tenslotte komen de strategieën voor verbetering van de duurzaamheid aan bod die in de literatuur genoemd worden, met een aanbeveling voor de beste strategie, zo die bestaat.

Hoofdstuk 2

Duurzame digitale informatie

Met de groei van de digitale informatie groeide ook het besef dat digitale informatie van een heel andere orde was dan de traditionele vormen van informatiedragers als papier, muziek- en videoband. Allerhande initiatieven werden ontplooid om het internet, tijdschriften, kranten en allerhande andere digitale informatie veilig te stellen voor latere generaties. Deze strategieën zijn onderwerp van hoofdstuk drie, voor die besproken kunnen worden moet eerst duidelijk zijn wat de problemen zijn van digitale informatie. Ook is het van belang duidelijk te weten wat duurzame digitale informatie precies is en hoe het in dit document gebruikt wordt, dat is onderwerp van dit hoofdstuk.

Het woord informatie wordt veelvuldig gebruikt in het informatietijdperk waarin we nu leven, wat het precies betekent blijft vaak onduidelijk, mede door het ambigue gebruik van het woord informatie¹. De vraag naar de betekenis van informatie wordt hier gelaten voor wat die is, dat is een taak voor filosofen. De UNESCO² ziet ons culturele erfgoed als informatie³. Digitale informatie is dan ons culturele erfgoed in digitale vorm, dat wil zeggen, informatie op een of andere manier gecodeerd in bits.

Deze digitale informatie bestaat uit documenten van allerlei aard, van kale tekst tot multimediale presentaties. Echter een document is zelf weer digitale informatie. In dit document zullen de termen ‘digitale informatie’ en ‘document’ daarom door elkaar gebruikt worden. Het gaat te ver om hier dieper in te gaan op de verschillende documentsoorten en bijbehorende bestandsformaten, enkel de opmerking dat er open en gesloten bestandformaten zijn, is van belang voor du-

¹Dretske, Fred I., *Knowledge and the flow of information* (Cambridge Mass 1999; 1e druk 1981) viii.

²United Nations Educational, Scientific and Cultural Organization (www.unesco.org).

³Lusenet, Yola de, ‘Preservation of digital heritage’ (<http://www.knaw.nl/ecpa/PUBL/-unesco.html>).

urzaamheid. Immers gesloten formaten zijn enkel duurzaam als daar commerciële motieven voor zijn.

Een informatiesysteem wordt gezien als een systeem dat digitale informatie, oftewel documenten, toegankelijk maakt. De nadruk zal liggen op het webinformatiesystemen, dat wil zeggen op informatiesystemen die via het internet documenten toegankelijk maakt.

duurzame digitale informatie wordt hier gedefinieerd als digitale informatie die enkele generaties goed toegankelijk blijft met minimaal verlies van informatie, in welke zin dan ook. Dus buiten het behoud van inhoud zijn ook het uiterlijk, functionaliteit, metadata, integriteit, authenticiteit, enz., van belang. Natuurlijk is het streven van duurzaamheid om de levensduur te maximaliseren, maar enkele generaties is een stuk realistische dan een oneindige levensduur.

Duurzame informatiesystemen tot slot, zijn informatiesystemen waarvan de documenten duurzaam gemaakt kunnen worden. Oftewel informatiesystemen die het duurzaam zijn of worden van digitale informatie zo goed mogelijk ondersteunen. Een lange levensduur voor de informatiesystemen an sich is zo goed als onmogelijk, de hard- en software veranderen en verouderen met een zo'n hoog tempo dat een levensduur van enkele decennia uitzonderlijk genoemd mag worden. De duurzaamheid van een informatiesysteem valt en staat dus met het overleven van de informatie in het systeem.

Hoofdstuk 3

Het probleem van duurzame digitale informatie

Het belang van duurzame digitale informatie is hetzelfde als het belang van gewone informatie: het behouden van ons culturele erfgoed voor het nageslacht. Door het vluchtige karakter van digitale informatie zijn er een aantal problemen die het belang van duurzame digitale informatie enkel onderstrepen. Deze problemen vallen uiteen in vier groepen: praktisch, betrouwbaarheid, selectie en context.

3.1 Een praktisch probleem: hard- en software

Digitale informatie is zoals eerder opgemerkt vluchtig. Het is een rij van bits, een rij van nullen en enen die zonder de juiste hardware en software niet te interpreteren is. Dat is aan de ene kant een sterk punt van digitale informatie, kopiëren is namelijk erg eenvoudig en betrouwbaar. De meest ingewikkelde documenten als afbeeldingen, muziek of video is in de basis even eenvoudig als een simpel tekstdocument, namelijk twee waarden, nul of een. Het kopiëren van niet-digitale informatie, bijvoorbeeld ene boek of een film gaat nooit zonder informatieverlies. Nadeel is wel dat de toegankelijkheid van digitale informatie volledig afhankelijk is van de juiste hardware en software, van de juiste interpretatie van de bits. Zelfs het meest eenvoudige karakter bestaat al uit acht bits, dus tweehonderdzesenvijftig mogelijke waarden per byte¹. Een iets minder Spartaans formaat kent al vele, vele malen meer mogelijkheden, laat staan videoformaten.

Voorals de hardware is een probleem omdat de verschillende architecturen en

¹Tenminste als we aannemen dat de eenvoudigste karakterencodering de ASCII (American Standard Code for Information Interchange) is. De nieuwere Unicode codering kent zestien bits codes, en kan dus 65536 tekens coderen.

apparatuur elkaar zo snel opvolgen, gemiddeld binnen twee tot vijf jaar². Dit is een probleem omdat voor elke nieuwe versie van hardware ook weer andere software nodig is. Trouwens de levensduur van software is niet veel langer dan die van de hardware. Buiten dat, datadragers worden zo goed als onbruikbaar als er geen leesapparaten meer zijn. Zijn er nog wel leesapparaten, dan is de vraag of die wel met moderne computers of software willen werken, er is zo al snel sprake van een compatibiliteitsprobleem³.

De levensduur van de meeste datadragers was voor de optische schijf erg kort, denk aan magnetische banden of acht inch floppy's⁴, maar ook gewone floppy's gaan vaak genoeg kapot binnen enkele jaren. In elk geval zijn zulke media niet duurzaam te noemen. Zelfs van optische schijven is een lange levensduur niet te garanderen⁵, dus om duurzaamheid te bereiken zou bij elke grote verandering in hardware alle informatie gekopieerd moeten worden op nieuwe datadragers, dat is echter wel een erg tijdsintensieve bezigheid, maar er valt niet aan te ontkomen.

Juist voor het bewaren van informatie van het internet zijn datadragers van levensbelang, het internet heeft namelijk geen geheugen. Het bestaat uit een grote verzameling van aan elkaar gekoppelde machines die elk apart verantwoordelijk zijn voor het behoud van de informatie. Door middel van hyperlinks worden allerhande documenten aan elkaar gekoppeld, maar die koppeling kan niet gegarandeerd worden voor meer dan enkele jaren. De kans dat een hyperlink uit een document na enkele decennia nog bruikbaar is, is nul. Dus om hypermediadocumenten, een groot deel van het tegenwoordige internet dus, duurzaam te kunnen bewaren, zijn er datadragers nodig.

3.2 Is digitale informatie wel te vertrouwen?

Digitale informatie is erg eenvoudig te kopiëren en te veranderen, leuk voor grappige familiefoto's, maar een probleem voor de betrouwbaarheid van digitale informatie. Immers veel delen van de maatschappij zijn afhankelijk van controleerbaarheid, denk aan de journalistiek of de wetenschap. Volgens David Bearman en Jennifer Trant is het zelfs erg waarschijnlijk dat er vele kopieën van een document in de omloop zijn waarvan er enkele niet overeenkomen met het origineel⁶.

²Waters, Donald en John Garrett, *Preserving digital information report of the task force on archiving of digital information* (Washington 1996) 4.

³Lusenet, Yola de, 'Preservation of digital heritage' (<http://www.knaw.nl/ecpa/PUBL/-unesco.html>).

⁴Rothenberg, Jeff, *Avoiding technological quicksand. Finding a viable technical foundation for digital preservation* (Amsterdam 1999) 7.

⁵Ibidem, 7

⁶Bearman, David en Jennifer Trant, 'Authenticity of digital resources. Towards a statement of requirements in the research process' (<http://www.dlib.org/dlib/june98/06bearman.html>).

Controleerbaarheid wordt dan zo goed als onmogelijk, duurzaamheid van die informatie daarmee nihil, immers oncontroleerbare informatie is niet bruikbaar.

Trouwens die controleerbaarheid is op het internet sowieso bijna onmogelijk, de levensverwachting van een webpagina ligt namelijk tussen vierenveertig dagen en twee jaar⁷, de kans dat een van de referenties in dit document ‘dood’ is, is dus erg groot. Daarbij is het internet een verzamelplaats van allerhande informatie, van universiteiten tot splintergroeperingen van vreemde sekten. Waar vroeger redacties van tijdschriften en kranten, bibliotheken en andere kennisinstellingen aangaven welke informatie autoriteit bezat, daar is nu het wereldwijde web waar alles en iedereen mag publiceren, dus de betrouwbaarheid van de informatie is moeilijker te bepalen.

Dat is nu al een probleem, want citeren of gebruiken van bronnen waarvan de auteur geen autoriteit bezit, levert onbetrouwbare informatie op. Nu kun je meestal wel achterhalen of een bron te vertrouwen is, maar hoe dat te doen als de hyperlinks niet meer bestaan?

3.3 Wie wat bewaart, heeft wat

Zoals hierboven gezegd, was voorheen de selectie van materiaal om te bewaren in handen van de welbekende instellingen als bibliotheken en archieven. Het probleem met digitale informatie is echter dat het niet duidelijk is wat nu precies bewaard moet worden. Off-line informatiesystemen, op cdrom bijvoorbeeld, kunnen eenvoudig in een kast gestopt en gecatalogiseerd worden, rest enkel het hardwareprobleem. Voor het internet daarentegen, het selecteren is een groter probleem: moet alles bewaard of enkel een select gedeelte, en wat te doen met afhankelijkheden⁸. Of wat dat betreft wanneer: webpagina’s hebben als eerder gezegd een korte levensduur, en veranderen bovendien regelmatig, laat staan de zogenaamde dynamisch gegenereerde webpagina’s⁹: hoe kom je daaraan?

Bijkomend probleem van het bewaren van het internet is het te pakken krijgen van de informatie. Vaak wordt het automatisch ophalen van informatie geblokkeerd of beperkt, ook het feit dat veel informatie enkel op aanvraag te verkrijgen is¹⁰ (bijvoorbeeld door een formulier in te vullen of bepaalde andere menselijke

⁷Kenney, Anne R., et al, ‘Preservation risk management for web resources’ (<http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january02/kenney/01kenney.html>).

⁸Lusenet, Yola de, ‘Preservation of digital heritage’ (<http://www.knaw.nl/ecpa/PUBL/enesco.html>).

⁹Kenney, Anne R., et al, ‘Preservation risk management for web resources’ (<http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january02/kenney/01kenney.html>).

¹⁰Arvidson, Allan, Krister Persson en Johan Mannerheim, ‘The Kulturarw project - the royal swedish web archive - an example of ”complete collection of web pages”’ (<http://www.ifla.org/IV/ifla66/papers/154-157e.htm>).

handelingen) uit te voeren, is het compleet bewaren van een beperkt deel van het internet al zo goed als onmogelijk.

Daarbij komt het dat er steeds meer instellingen, personen en instanties zelf publiceren in plaats dat door uitgeverijen of tijdschriften te laten doen. Was vroeger een artikel in een wetenschappelijk tijdschrift betrouwbaar te noemen omdat er een professionele redactie deze ‘goed’ beoordeeld had, nu moet eenieder dat zelf uitzoeken als op een website een wetenschappelijk document gevonden word. Hier zou een taak kunnen liggen voor overheden of bibliotheken, maar dan wordt het wel mensenwerk, automatiseren is door de belabberd gebruik van metadata op internet niet mogelijk¹¹.

Een ander punt zijn de uitgeverijen, voornamelijk van tijdschriften en kranten. Ook zij zijn overgestapt op digitale informatie, en veel bibliotheken hebben nu niet de echte tijdschriften op papier in bezit, maar een licentie om de informatie bij de uitgevers op te vragen. Ook zaken als copyrights beperken de mogelijkheid om informatie te bewaren door onafhankelijke instellingen¹². Hierdoor komt een deel van het culturele erfgoed in handen van commerciële organisaties. De beschikbaarheid daarvan is dan niet meer gegarandeerd.

3.4 Informatie is meer dan enkel inhoud

Een ander probleem is de authenticiteit van digitale informatie, immers digitale informatie is snel omgezet van het ene formaat naar het andere. Een verandering is snel doorgevoerd. Ook de context van het verschijnen van de informatie verandert regelmatig. Waar vroeger de informatie naast de inhoud nog een duidelijke context had, daar is nu enkel de inhoud. Een digitale krant biedt veel minder informatie over informatie dan een krant van papier. Immers een artikel op een voorpagina is van groter belang dan een op de laatste pagina, in kleine lettertjes. Maar een digitale voorpagina bestaat hoogstens een dag, en daarna is het artikel enkel te zien in een andere context, bijvoorbeeld een zoekmachine.

Ook opmaak en structurering bieden extra informatie over de inhoud van een document. Door accentuering kan een auteur beeldspraak toepassen, of een voorbeeld kan zo onderscheiden worden van gewone tekst. Wat te denken van tabellen? Dra er bij een omzetting van het ene formaat in het andere iets fout gaat met de tabel, kan het desastreuze gevolgen hebben. Zo zijn er meer kleinigheden te bedenken van informatie die niet zozeer inhoudelijk doch toevoegend zijn.

¹¹Day, Michael, ‘Resource discovery, interoperability and digital preservation. Some aspects of current metadata research and development’, (<http://www.ukoln.ac.uk/metadata/publications/vine-117/>).

¹²Lusenet, Yola de, ‘Preservation of digital heritage’ (<http://www.knaw.nl/ecpa/PUBL/-enesco.html>).

Vooral voor historisch onderzoek zijn deze kleinigheden van groot belang voor de 'juiste' interpretatie van een bron.

Hoofdstuk 4

Strategieën ter vergroting der duurzaamheid

In het vorige hoofdstuk is besproken welke problemen er zijn van duurzame digitale documenten. Voor sommige problemen zijn in de literatuur enkele strategieën geopperd om deze te verminderen. Opvallend is wel dat deze strategieën voornamelijk bij archief- en bibliotheekwetenschappers te vinden zijn. Informatici schijnen er niet over na te denken, dat terwijl het beter zou zijn als de makers en diegenen die de digitale informatie bewaren samen een strategie zouden bedenken en uitvoeren¹. Immers het maken van digitale informatie nu zegt niets over hoe het later gebruikt gaat worden². Wat nu voor de hand liggend moge zijn zou over enkele jaren wel eens belangrijke informatie kunnen zijn om de documenten van nu te interpreteren.

4.1 Afgedrukt en wel, maar niet digitaal

Al eeuwenlang is papier de informatiedrager bij uitstek. De maatschappij is gewend er mee om te gaan. Het bewaren, het duurzaam beschikbaar houden van informatie op papier is geen groot probleem. Zorgen voor goede omstandigheden voor het papier en het gaat eeuwen mee, rampen en ongelukken uitgezonderd. Bijkomend voordeel van papier is dat voor dit medium zeer goede leesapparaten zijn: de mens. Talen mogen uit de gratie raken, maar het kan geleerd worden omdat het aan begrijpelijke eenvoudige regels voldoet. Na wat onderzoek kunnen de meeste teksten in een oude taal gelezen worden.

¹Lusenet, Yola de, 'Preservation of digital heritage' (<http://www.knaw.nl/ecpa/PUBL/-enesco.html>).

²Day, Michael, 'Metadata for digital preservation: an update', (<http://www.ariadne.ac.uk/-issue22/metadata/>).

Digitale informatie is op de basis nog eenvoudiger, er zijn maar twee mogelijke waarden, maar hierdoor is de informatie erg complex gecodeerd. Het is dan ook niet vreemd dat er geopperd is om daarom digitale informatie af te drukken op papier en zo die informatie te behouden³. Dit is echter geen goede oplossing: eenvoudige digitale documenten moge dan afdrukbaar zijn, documenten die de digitale mogelijkheden uitbuiten, kunnen niet op een zinvolle manier geprint worden⁴. Moderne digitale informatie bestaat uit hyperlinks, multimedia, of wordt gegenereerd uit een database, probeer dat maar eens af te drukken. Aan de andere kant wordt er nog veel documenten gemaakt die enkel digitaal bestaan doch bedoeld zijn voor papier zoals veel artikelen, deze kunnen goed afgedrukt worden. In feite gebeurt dat zo ook, tijdschriften kunnen nu op papier of via het internet geraadpleegd worden.

4.2 Oudheidkundige informatica: computermusea

Naast afdrukken is nog een andere vreemde oplossing aangedragen: computermusea. In deze musea zouden oude computers en de daarvoor beschikbare software bruikbaar gehouden moeten worden zodat oude digitale informatie altijd toegankelijk blijft⁵. Het is wel een leuke oplossing, en computermusea moeten er zeker komen, computers en hun software horen namelijk ook tot het culturele erfgoed, maar of het een werkbare oplossing is?

Hardware en veel datadragers hebben nu immers niet het oneindige leven, eerder een korte levensduur⁶. De kans dat over honderd jaar onze machines nog zullen werken lijkt me onwaarschijnlijk, laat staan dat de datadragers nog leesbaar zijn. Die informatie op die datadragers moet dan toch overgezet worden op nieuwe datadragers, en die kunnen weer niet op de oude machines gelezen worden met oude software en apparaten, nieuwe software en apparaten kunnen niet aangesloten worden op oude machines.

Een bijkomend probleem is dat hardware en software erg snel vervangen worden door nieuwe versies. Om al die verschillende versies in een museum te bewaren is op de lange termijn ondoenlijk. Om maar niet te spreken van de administratie en kennis die nodig is om al die verschillende datadragers, bestandsformaten, machines, software en apparaten met elkaar te laten werken. Immers het ene bestandsformaat kan enkel op die machine, die datadrager in dat apparaat, maar die moet dan weer die software hebben, enz. Kortom deze weg is een doodlopende.

³Rothenberg, Jeff, *Avoiding technological quicksand. Finding a viable technical foundation for digital preservation* (Amsterdam 1999) 3.

⁴Ibidem, 9

⁵Ibidem, 12

⁶Ibidem, 13

Dan is daar het probleem van het internet: hoe deze te bewaren? Het web is een dynamisch lichaam, die met de seconde verandert en groeit. Probeer dit internet maar te bewaren, en behoud de illusie dat het een authentieke weergave van het origineel is dat in de tijd is geëvolueerd is.

4.3 Na het een komt het ander: migratie

Een andere strategie is die van migratie van informatie van het ene formaat naar een nieuwer formaat. Dra een formaat uit de gratie geraakt moet dan alle informatie omgezet worden in een nieuwer formaat. Deze strategie is de meest gebruikte strategie, en kent een langere geschiedenis doordat paradigmaveranderingen en formaatswijzingen in de informatica aan de orde van de dag zijn⁷. Rothenberg is een duidelijk tegenstander van deze strategie, als blijkt in zijn *Avoiding technological quicksand*: ‘het is beter dan niets’⁸ vindt hij. Grote struikelblokken zijn volgens hem het feit dat migratie arbeidsintensief is, veel tijd kost, foutgevoelig is, en dat er altijd de kans bestaat informatie te verliezen, voornamelijk omdat paradigmaverschuivingen verrassende wendingen nemen zodat elke migratie nieuwe problemen met zich mee brengt⁹. Volgens Bearman slaat Rothenberg de plank volledig mis omdat hij het over migratie van informatiesystemen heeft in plaats van migratie van documenten, en gebruikt daarbij enkel twee bronnen¹⁰, en betwijfeld daarmee de autoriteit van dat stuk.

Michael Day sluit zich aan bij Bearman met kritiek op Rothenberg, maar ziet bij migratie wel een groot probleem doordat migratie informatieverlies, verlies van functionaliteit, bruikbaarheid en integriteit tot gevolg zal hebben¹¹. Hij draagt daarvoor een oplossing aan: beter gebruik van metadata¹², zijn paradepaardje.

Een ander probleem dat Rothenberg aansnijdt is het moeilijk automatiseren van migraties¹³, maar dat probleem kan grotendeels opgelost worden door het gebruik van goed gedefinieerde en gestructureerde standaarden als SGML¹⁴ of XML¹⁵. Immers zulke documenten in zulke standaarden kunnen door al dan niet

⁷Ibidem, 13

⁸Ibidem, 16

⁹Ibidem, 13

¹⁰Bearman, David, ‘Reality and chimeras in the preservation of electronic records’ (<http://www.dlib.org/dlib/april99/bearman/04bearman.html>).

¹¹Day, Michael, ‘Metadata for digital preservation: an update’, (<http://www.ariadne.ac.uk/issue22/metadata/>).

¹²Ibidem

¹³Rothenberg, Jeff, *Avoiding technological quicksand. Finding a viable technical foundation for digital preservation* (Amsterdam 1999) 15.

¹⁴Standard Generalized Markup Language

¹⁵eXtensible Markup Language

eenvoudige regels getransformeerd worden in andere gestructureerde formaten, maar over standaarden in de volgende paragraaf meer.

4.4 Waren standaarden maar standaard ...

Standaarden zijn er legio, en worden door alle gebieden der informatica gebruikt, ook in informatiesystemen, bestandsformaten en het internet, denk aan SQL¹⁶ of HTML¹⁷. Er bestaat een belangrijk onderscheid tussen standaarden: openheid. Vooral bedrijven hebben in het verleden gesloten standaarden gebruikt om commerciële redenen. Duidelijk is nu dat dit de duurzaamheid niet ten goede is gekomen. Zodra een onderneming een standaard niet meer kan of wil ondersteunen, is alle informatie die volgens die standaard is opgezet onleesbaar geworden.

Open standaarden daarentegen is een beter idee: de definitie van de standaard is vrijelijk beschikbaar, en daardoor blijven documenten in een open standaard langer leesbaar. Zelfs als de standaard in onbruik is geraakt, en dat gebeurt meestal al in enkele softwaregeneraties¹⁸ dan kan, indien er een definitie ervan bewaard is gebleven, een nieuwe implementatie gemaakt om informatie volgens zo een standaard te interpreteren. Nadeel van standaarden verder is dat zij erg snel veranderen¹⁹, dus migratie zal al snel nodig zijn. Voordeel is wel dat bij gestructureerde standaarden die migratie automatisch verlopen kan, blijft enkel over het probleem van informatieverlies, hoe klein ook.

Echter het gebruik van open standaarden krijgt enkel medewerking van de industrie als er commerciële argumenten voor zijn²⁰, en zulke argumenten zijn er enkel als de markt er om vraagt. Toch valt of staat de duurzaamheid door standaarden door acceptatie ervan door de industrie²¹, dus een vorm van overheidsinmenging is hier op zijn plaats. Gelukkig zijn er organisaties als het World Wide Web Consortium²² die standaarden definiëren, onderhouden en propageren.

¹⁶Structured Query Language

¹⁷HyperText Markup Language

¹⁸Bearman, David, 'Reality and chimeras in the preservation of electronic records' (<http://www.dlib.org/dlib/april99/bearman/04bearman.html>).

¹⁹Rothenberg, Jeff, *Avoiding technological quicksand. Finding a viable technical foundation for digital preservation* (Amsterdam 1999) .

²⁰Bide, Mark, 'Standards for electronic publishing' (Amsterdam 2000, elektronisch beschikbare versie gebruikt: url nog opzoeken).

²¹Lusenet, Yola de, 'Preservation of digital heritage' (<http://www.knaw.nl/ecpa/PUBL/-unesco.html>).

²²<http://www.w3c.org>

4.5 Informatie over informatie

Metadata is informatie over informatie, zeg maar dat het de beschrijving van de context van informatie is. Het gaat dan om gegevens als auteur, uitgever, plaats en dergelijke evidente zaken, maar ook over uiterlijkheden, gebruikte software, hardware, of over de inhoud, zoals waarover het gaat, trefwoorden en dergelijke. Zulke informatie is belangrijk voor duurzaamheid van informatie omdat zo de informatie toegankelijk blijft, niet enkel in inhoudelijke zin, maar ook van de context.

Dra deze metadata gestandaardiseerd wordt, is het zelfs mogelijk automatisch te catalogiseren²³. Zulke standaardisatie is trouwens nodig want om het internet te kunnen behouden of zelfs stukken ervan is het nodig om de bestaande metadata sterk te verbeteren, nu worden grote delen van het internet namelijk al niet eens gevonden door de catalogi als zoekmachines²⁴. En om informatie te kunnen behouden, moet het wel eerst gevonden kunnen worden.

In elk geval is metadata zeer bruikbaar bij alle andere strategieën, die kunnen niet zonder extra informatie over de te bewaren informatie²⁵. Bij het afdrukken kan het gebruikt worden op de normale wijze in bibliotheken. Voor musea kan het ook als catalogusinformatie en administratie dienen. Bij migratie is het nodig om alle informatie buiten de inhoud te kunnen bewaren, om maar het origineel te kunnen blijven benaderen bij elke migratie. Emulatie gebruikt metadata om informatie over de hardware en software te verzamelen. Ook informatie over gebruikte standaarden is zeer bruikbaar, immers een gekende standaard maakt het decoderen van digitale informatie een stuk eenvoudiger.

Tegenwoordig wordt door uitgevers al wel bibliografische data aangeleverd bij hun producten, enkel deze informatie laat erg te wensen over²⁶, dus is er wel een of andere vorm van standaardisatie nodig. En zijn er, denk aan RDF²⁷ dat steeds meer bekendheid krijgt.

²³Lusenet, Yola de, 'Preservation of digital heritage' (<http://www.knaw.nl/ecpa/PUBL/-unesco.html>).

²⁴Day, Michael, 'Resource discovery, interoperability and digital preservation. Some aspects of current metadata research and development', (<http://www.ukoln.ac.uk/metadata/-publications/vine-117/>).

²⁵Day, Michael, 'Extending metadata for digital preservation', (<http://www.ariadne.ac.uk/-issue9/metadata/>).

²⁶Bide, Mark, 'Standards for electronic publishing' (Amsterdam 2000, elektronisch beschikbare versie gebruikt: url nog opzoeken).

²⁷Resource Description Framework

4.6 Emulatie

De laatste strategie die in de literatuur besproken wordt is emulatie. Emulatie is niets meer dan het idee om een echte machine in software volledig te simuleren en op een andere echte machine te laten draaien. Het idee is dus om oude computers in software na te bouwen en daarop de oude software te draaien, zodat de informatie er precies het zelfde zal blijven uitzien.

Vooraf Jeff Rotenberg is een groot voorstander van deze strategie, in zijn ‘Avoiding technological quicksand’ gaat hij er uitvoerig op in nadat hij de andere strategieën heeft afgewezen. In het kort komt zijn oplossing er op neer dat er algemene technieken voor emulatoren gevonden moeten worden, dat de juiste metadata voor bruikbaarheid gevonden moeten en dat er technieken om de informatie, software, hardware-emulator en metadata te encapsuleren zodat er een algemene manier met oude informatie omgegaan kan worden²⁸.

Het is een complexe strategie en volgens Bearman is het overdreven complex omdat de informatiesystemen worden bewaard in plaats van de ‘records’²⁹. Lusenet ziet deze oplossing enkel op korte termijn van nut omdat met den duur er dan emulaties van emulaties van emulaties komen die onhandelbaar zullen zijn³⁰. Trouwens het is maar de vraag of volledige emulatie technisch wel zal werken³¹, is het wel mogelijk om het origineel volledig te benaderen? Zo niet dan is deze oplossing een te omslachtige om te bereiken wat ook eenvoudiger kan.

4.7 En de beste strategie is ... Een aanbeveling

De vraag wat de beste strategie is, is zo goed als niet te beantwoorden. Wel is duidelijk dat het afdrukken en de computermusea geen echt werkbare oplossingen zijn, hoewel ze op de korte termijn en bepaalde deelgebieden zeker zullen kunnen werken. Ook de emulatie strategie is erg moeilijk te handhaven als een oplossing, ze is te complex.

Wat zeker belangrijk zal zijn is goede gestructureerde metadata en open standaarden zodat automatische migratie mogelijk is. De problemen van de datadragers blijven natuurlijk bestaan, toch zijn deze drie gecombineerd de beste strategie voor duurzaamheid.

²⁸Rothenberg, Jeff, *Avoiding technological quicksand. Finding a viable technical foundation for digital preservation* (Amsterdam 1999) 17.

²⁹Bearman, David, ‘Reality and chimeras in the preservation of electronic records’ (<http://www.dlib.org/dlib/april99/bearman/04bearman.html>).

³⁰Lusenet, Yola de, ‘Preservation of digital heritage’ (<http://www.knaw.nl/ecpa/PUBL/enesco.html>).

³¹Bearman, David, ‘Reality and chimeras in the preservation of electronic records’ (<http://www.dlib.org/dlib/april99/bearman/04bearman.html>).

Een probleem lost ze niet op, het feit dat hyperlinks dood gaan na verloop van tijd, net als literatuurverwijzingen. Daartegen is geen metadata gewassen, de externe documenten zullen, indien mogelijk, bij het te bewaren document gevoegd moeten worden. Natuurlijk is dat voor video en muziek een probleem in verband met de beschikbare opslagruimte. Tevens is er het risico van cascaderende of recursieve toevoegingen van externe documenten. Neemt niet weg dat als ook dit toevoegen van externe documenten gestandaardiseerd kan worden op een gestructureerde manier met voldoende metadata het een strategie zou kunnen zijn die het beste toekomst heeft. Of dat echt zo is, daarvoor is extra onderzoek nodig.

Hoofdstuk 5

Conclusie

De centrale vraag in dit document is hoe het gesteld staat met de duurzaamheid van informatiesystemen. Allereerst werden voor de termen duurzame digitale informatie en duurzame informatiesystemen definities gegeven zoals ze in dit document gebruikt worden.

Duurzame digitale informatie zijn documenten die enkele generaties toegankelijk blijven met een minimaal verlies aan informatie. Duurzame informatiesystemen zijn systemen die duurzame digitale informatie mogelijk maken, maar zullen zelf niet duurzaam zijn, de documenten zullen hun originele informatiesystemen verre overleven.

Nadat duidelijk is wat duurzame digitale is, werden de problemen die er zijn om die duurzaamheid te verkrijgen onder de loep genomen. Deze problemen vallen in vieren uiteen. De levensduur van hard- en software maar ook van datadragers zijn veel te kort om te verwachten dat digitale informatie behouden blijft in de originele omgeving.

Daarnaast is het internet een enorme vergaarbak van informatie, waar niet duidelijk is welke documenten nu autoriteit bezitten en welke niet. Er zal, mede door het grote aantal documenten, geselecteerd moeten worden, voor het bewaren van documenten, een taak voor bibliotheken. Maar wat moet nu wel en niet geselecteerd? Alles is in elk geval niet mogelijk, dus keuzen moeten gemaakt.

Het laatste probleem is het probleem van informatie naast de inhoud: de context. De context is erg moeilijk te bewaren doordat het internet een dynamisch lichaam is, verhoudingen van documenten ten opzichte van elkaar veranderen voortdurend, en zijn dus snel verdwenen.

Na de problemen zijn de in de literatuur voorkomende strategieën ter oplossing ervan besproken. Twee ervan, printen van digitaal materiaal en computermusea, zijn niet bruikbaar. Het eerste omdat daardoor het digitale aspect verloren gaat, als het al uitvoerbaar is, hoe print je een hypermediadocument? De tweede oplossing is onbetrouwbaar, hardware doorstaat de tand des tijds nu eenmaal

niet.

Een complexe oplossing, gepropageerd door Rothenberg, emulatie van oude hardware en daarop de oude software draaien en die gegevens in combinatie met de informatie, metadata en al te bewaren is niet bruikbaar omdat het niet zozeer documenten duurzaam wil maken als wel de informatiesystemen. Trouwens het is maar de vraag of emulatie wel technisch mogelijk is.

Enkele meer belovende strategieën zijn metadata, standaarden en migratie, en vooral gecombineerd zijn ze krachtig. open standaarden gebruiken, bevordert de toegankelijkheid op de lange termijn omdat de specificaties bekend zijn en altijd opnieuw weer te geïmplementeerd kunnen worden. Metadata is informatie over informatie, nodig om de informatie naast de inhoud te bewaren. Migratie is het omzetten van een document in het ene formaat naar een ander. Indien die formaten gestructureerde open standaarden zijn en als nu ook de metadata gestructureerd en zo volledig mogelijk is, kan migratie zo goed als automatisch.

Het enige probleem is dan het bewaren van externe documenten bij een document, maar dat moet wel, vertrouwen op het geheugen van het internet is geen goed idee, immers het internet leidt, hoe jong het ook is, aan Alzheimer. Een document blijft hoogstens enkele jaren bestaan op het internet.

Literatuur¹

Arvidson, Allan, Krister Persson en Johan Mannerheim, 'The Kulturarw project - the royal swedish web archive - an example of "complete collection of web pages" (<http://www.ifla.org/IV/ifla66/papers/154-157e.htm>).

Bearman, David en Jennifer Trant, 'Authenticity of digital resources. Towards a statement of requirements in the research process' (<http://www.dlib.org/dlib/june98/06bearman.html>).

Bearman, David, 'Reality and chimeras in the preservation of electronic records' (<http://www.dlib.org/dlib/april99/bearman/04bearman.html>).

Bide, Mark, 'Standards for electronic publishing' (Amsterdam 2000, elektronisch beschikbare versie gebruikt: url nog opzoeken).

Day, Michael, 'Extending metadata for digital preservation', (<http://www.ariadne.ac.uk/issue9/metadata/>).

Day, Michael, 'Metadata for digital preservation: an update', (<http://www.ariadne.ac.uk/issue22/metadata/>).

Day, Michael, 'Resource discovery, interoperability and digital preservation. Some aspects of current metadata research and development', (<http://www.ukoln.ac.uk/metadata/publications/vine-117/>).

Dretske, Fred I., *Knowledge and the flow of information* (Cambridge Mass 1999; 1e druk 1981).

Hedstrom, Margaret, 'Research challenges in digital archiving and long-term preservation' (http://www.sis.pit.edu/dlwkshop/paper_hedstrom.pdf).

Holdsworth, David en Paul Wheatley, 'Emulation, preservation and abstraction' (<http://129.11.152.25/CAMiLEON/dh/ep5.html>).

Kenney, Anne R., et al, 'Preservation risk management for web resources' (<http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january02/kenney/01kenney.html>).

¹Bij de annotatie door heel het document worden de url's die te lang zijn om op een regel te passen, afgebroken na een '/' door middel van een koppelteken, dat koppelteken behoort dus niet tot de url.

- Lupovivi, Catherine en Julien Masanès, *metadata for the long term preservation of electronic publications* (Amsterdam 2000).
- Lusenet, Yola de, 'Preservation of digital heritage' (<http://www.knaw.nl/-ecpa/PUBL/enesco.html>).
- Rothenberg, Jeff, *Avoiding technological quicksand. Finding a viable technical foundation for digital preservation* (Amsterdam 1999).
- Waters, Donald en John Garrett, *Preserving digital information report of the task force on archiving of digital information* (Washington 1996).