ARCHIVEREN VAN DIGITAAL ACADEMISCH ERFGOED

Archiveren van Digitaal Academisch Erfgoed

Een verslag als voorbeeld

Onder redactie van
Heiko Tjalsma

DANS studies in digital archiving 2
Den Haag, 2006

DANS - Data Archiving and Networked Services
Postbus 93067
2509 AB Den Haag
T 070 3494450
F 070 3494451
info@dans.knaw.nl
www.dans.knaw.nl

ISBN 90-6984-501-6

Het papier van deze publicatie voldoet aan ∞ iso-norm 9706 (1994) voor permanent houdbaar papier.

Druk: Krips
Ontwerp en dtp: Ellen Bouma, Edita-KNAW
Productie: Edita-KNAW

Illustratie voorkant: Photograph courtesy of the School of Computing and IT at the University of Wolverhampton. We thank Professor Robert Moreton for giving permission to reproduce the photograph in this publication.
The picture shows the institution's earliest computer called the WITCH which stood for Wolverhampton Instrument for Teaching Computing from Harwell. It was built at AERE Harwell in 1948 and was won by the then Wolverhampton and South Staffordshire Technical College in a national competition in 1957. The WITCH was a very slow computer by modern standards. It took 2 seconds to add or subtract 2 numbers, 5 seconds to multiply two numbers and 15 seconds to divide two numbers. (Division by zero took rather longer.) The photograph was taken in 1961.

Alle in deze publicatie genoemde URL's zijn geconsulteerd in juli 2006

Woord vooraf

Van 2000 tot en met 2003 heeft het Nederlands Historisch Data Archief (NHDA) een project uitgevoerd om de aanpak, de mogelijkheden en de kosten te verkennen van het archiveren van bestaand maar niet geordend digitaal wetenschappelijk materiaal. Het project kreeg de naam ADA: Archiveren van Digitaal Academisch Erfgoed.

Het NHDA werd in 2005 onderdeel van DANS, Data Archiving and Networked Services, dat in dat jaar werd opgericht en expliciet werd belast met het bevorderen en faciliteren van de archivering van onderzoeksdata in Nederland. Terugkijkend kan dus worden vastgesteld dat het ADA-project een vingeroefening was voor een van de activiteiten die DANS inmiddels als dienst in de academische wereld aanbiedt onder de naam ADA: Academische Data Archivering. Vanuit dat perspectief is dit verslag geschreven. Het biedt een beknopte rapportage van het ADA-project en geeft tegelijk inzicht in de mogelijkheden van de nu door DANS aangeboden diensten.

Het oorspronkelijke project is gesubsidieerd vanuit het programma Innovatie Wetenschappelijke Informatievoorziening (iWI) van Stichting SURF en uitgevoerd op het Meertens Instituut. Het NHDA maakte ten tijde van het onderzoek deel uit van het NIWI, het Nederlands Instituut voor Wetenschappelijke Informatiediensten, een instituut van de Koninklijke Nederlandse Akademie van Wetenschappen (KNAW).

De werkzaamheden zijn voornamelijk verricht door projectmedewerker Tom van den Berg en projectleider Heiko Tjalsma. Daarnaast is er werk uitgevoerd door enkele andere toenmalige NIWI-medewerkers, in het bijzonder Richard Bos en Bram Buitendijk.

Vanuit het Meertens Instituut werd het project intensief begeleid door Koos Schell, terwijl assistentie werd verleend door haar collega's Carinqua van Wijk, Edwin Brinkhuis en Jan Pieter Kunst. Stagiair Ron Edel deed veel inventariserend werk.

Het project is begeleid door Peter Doorn, als hoofd van het NHDA destijds werkzaam bij het NIWI. Waardevolle adviezen kwamen ook van René van Horik, Marjan Balkestein en Annelies van Nispen (NHDA/NIWI), Frank Peeters (Afdeling Neerlandistiek/NIWI), Edo Dooijes (Computermuseum UvA) en Henk Voorbij (KB). Op eerdere versies van dit rapport is kritisch en deskundig commentaar

geleverd door René van Horik (NHDA/NIWI), door Cor van der Meer en Ruud Bronmans (beiden Steinmetzarchief/NIWI) en door Frans van der Kolff (NIWI).

Al deze personen verdienen dank voor hun welkome bijdrage. Dat geldt ook voor degenen die hebben meegewerkt aan het marktonderzoek en nog eens extra voor het Meertens Instituut, dat zijn data als proefveld beschikbaar stelde.

Van dit rapport zijn hoofdstuk 3 en de bijlagen B, C en D oorspronkelijk door Tom van den Berg geschreven, terwijl hoofdstuk 6 door Peter Doorn werd geleverd. De andere hoofdstukken en bijlage A (met bijdragen van Tom van den Berg) zijn geschreven door Heiko Tjalsma, die ook voor de eindredactie tekende. De definitieve versie van dit rapport is bewerkt door Martijn de Groot.

Heiko Tjalsma

Inhoud

1 Het ADA-project: achtergrond, doel en aanpak

1.1 Inleiding

Vrijwel de gehele Nederlandse wetenschappelijke productie is tegenwoordig digitaal. Hoezeer de academische wereld daaraan al gewend is, betrekkelijk nieuw is nog de vraag hoe het staat met de bewaring van die productie op lange termijn. Zijn de betrokken data en documenten over vijf of tien jaar nog toegankelijk en begrijpelijk?

De afgelopen jaren is het besef gegroeid dat ons digitale erfgoed in gevaar is. Het probleem geniet toenemende aandacht bij bibliotheken, overheidsarchieven en het bedrijfsleven, maar ook in de academische wereld. Het ADA-project Archiveren van Digitaal Academisch Erfgoed is uitgevoerd om een bijdrage te leveren aan de oplossing van deze problematiek, speciaal gericht op de onderzoekswereld.

1.2 Achtergrond van het onderzoek

In de huidige praktijk op het gebied van digitale archivering bestaat grote behoefte aan pilot-projecten om te experimenteren met langetermijnbewaring. De laatste jaren zijn bibliotheken en archieven op uiteenlopende schaal met zulke projecten begonnen.[1]

Er zijn echter verschillende archiveringsstrategieën die daarbij als uitgangspunt kunnen dienen. Uitersten zijn enerzijds emuleren en anderzijds migreren en converteren. Bij emulatie worden bestanden in het oorspronkelijke bestandsformaat bewaard en worden systemen ontwikkeld waardoor de originele software kan blijven functioneren. Bij migratie en conversie worden de bestanden omgezet naar formaten die door nieuwe software kan worden begrepen. Dit zijn, bij voorkeur, standaardformaten.[2] Bij het e-depot van de Koninklijke Bibliotheek (KB) wordt geëxperimenteerd met emulatie, maar de meeste data-archieven op het terrein van de alfa- en gammawetenschappen, waaronder die van DANS, maken ge-

1 Bijvoorbeeld het e-depot project van het Rotterdamse Gemeentearchief.

2 Voor een samenvatting zie Rothenberg (1999) en Bearman (1999).

bruik van migratie en conversie. Ook het ADA-project heeft zich vooral hierop geconcentreerd.[3]

Daarnaast wordt er verschillend gedacht over de plaats van bewaring. Kunnen bestanden beter door centrale, bij voorkeur landelijke, depots worden bewaard of moet dat juist decentraal gebeuren bij de instelling, die de bestanden heeft gecreëerd? Er is tot nu toe niet veel onderzoek verricht naar de vraag welke van deze twee opties de voorkeur verdient. Juist over dit aspect zou het ADA-project meer inzicht kunnen verschaffen.[4]

Ook maakt het bij de aanpak van het digitale bewaarprobleem verschil om wat voor type bestanden het gaat (tekstbestanden, databases, grafische bestanden), en door welk soort instelling ze bewaard worden. Er zijn verschillende functionaliteiten nodig voor de langetermijnbewaring van elektronische publicaties, digitale archiefstukken en digitale onderzoeksbestanden. Zo is het voor elektronische publicaties niet alleen van belang om de inhoud maar ook om de vorm zo goed mogelijk te bewaren, zeker in nationale depotbibliotheken met hun taak op het terrein van het nationale culturele erfgoed. Bij archiefstukken is juist de authenticiteit van het stuk van cruciaal belang vanwege de juridische status (bewijsvoering). Bij wetenschappelijke databestanden worden vorm en authenticiteit over het algemeen van minder groot belang geacht dan het gebruiksgemak voor secundaire analyse. Bij deze laatste groep van bestanden is echter vaak het ontbreken van regels voor het bewaren een reëel probleem, waardoor niemand zich uiteindelijk verantwoordelijk voelt.

Deze functionele verschillen vloeien voort uit verschillen in bewaarcontext: vanuit welke optiek wordt het bewaren als belangrijk gezien: een wetenschappelijke, administratieve of culturele? Verschillen in bewaarcontext maken op zichzelf nog geen afzonderlijke bewaarinstituten noodzakelijk. Integendeel: er zijn goede redenen om elektronische publicaties samen met de bijbehorende onderzoeksdata in dezelfde instelling, bijvoorbeeld een universiteit, te bewaren. Dat dit op het moment niet of nauwelijks gebeurt wordt primair veroorzaakt door beleidskeuzen, prioriteiten en vooral de historisch gegroeide werkterreinen van de onderscheiden instellingen.

Op dit terrein is er zodoende sprake van een belangrijke organisatorisch-institutionele factor, die aangeeft aan dat het digitale bewaringsprobleem bepaald niet alleen een technische dimensie heeft. Organisatorische kwesties spelen een minstens zo belangrijke rol. Eén van de conclusies van eerder uitgevoerd onderzoek[5] was dat in het bijzonder voor de bewaring van de digitale wetenschappelijke bronnen in Nederland op enkele uitzonderingen na geen goede voorzieningen aanwezig zijn. Zelfs van bewaarbeleid bleek vaak geen sprake te zijn. Het ADA-project

3 Voor het e-depot zie Van der Werf-Davelaar (2001). Recenter: Oltmans en Van Wijngaarden (2004).

4 Zie bijvoorbeeld Hedstrom (Londen 1995) of Doorn en Tjalsma (1997).

5 Mostert e.a. (1998).

richtte zich op het bewaren van dit onderzoeksmateriaal, dat wil zeggen van wetenschappelijke databestanden.

1.3 Doel van het ADA-project

Doel van het ADA-project was het beantwoorden van de vraag naar de haalbaarheid van het aanbieden van digitale archiveringsdiensten aan de wetenschappelijke wereld: universiteiten en onderzoeksinstituten, in eerste instantie op het gebied van de humaniora en de sociale wetenschappen. Dat is een andere activiteit dan wat tot nu toe gebruikelijk was bij de bestaande data-archieven, zoals het Steinmetzarchief of het NHDA. Onderzoeksbestanden werden bij het data-archief gedeponeerd door individuele onderzoekers, onderzoekgroepen, instituten en organisaties als het Sociaal en Cultureel Planbureau of het Centraal Bureau voor de Statistiek. Meestal werd voor deze deponering noch voor het verdere beheer en de ontsluiting door de rechthebbenden betaald.

Bij de digitale archiveringsdiensten waarop dit project zich richt kan aan verschillende mogelijkheden gedacht worden, in oplopende graad van intensiteit:

- het aanbieden van consultancy of cursussen op het gebied van technische, documentaire of organisatorische aspecten;
- het verzorgen van de fysieke opslag van de bestanden en assisteren bij de documentatie terwijl het beheer, met name de toegankelijkstelling, in handen blijft van de instelling;
- centrale archivering: de bestanden gaan volledig over in beheer van het centrale data-archief, inclusief de beschikbaarstelling.

Dit onderzoek moest informatie opleveren over de vraag op welke wijze zulke archiveringsdiensten uitvoerbaar zijn, of ze kostendekkend kunnen worden uitgevoerd en of wetenschappelijke organisaties erin geïnteresseerd zijn.

1.4 Opzet van dit verslag

Om vorengenoemde vragen te beantwoorden heeft om te beginnen een oriënterend onderzoek plaatsgevonden naar de nieuwste ontwikkelingen op het gebied van de langetermijnbewaring (hoofdstuk 2). Daarnaast is een marktonderzoek uitgevoerd, om na te gaan in hoeverre er in wetenschappelijk Nederland vraag is naar de bedoelde vormen van dienstverlening (hoofdstuk 4). Het belangrijkste onderdeel was echter het pilot-project 'Meertens Data' waarin de wetenschappelijke onderzoeksbestanden van het Meertens Instituut zijn gearchiveerd (hoofdstuk 3). Dit instituut doet 'etnologisch onderzoek van de functie, de betekenis en de onderlinge samenhang van cultuuruitingen alsmede taalstructureel, dialectologisch en sociolinguïstisch onderzoek naar taalvariatie binnen het Nederlands in Nederland, met een nadruk op grammaticale en naamkundige variatie'. Het koesterde al de wens een beleid te ontwikkelen voor het bewaren van vooral oudere onderzoeksbestanden. Daartoe zouden deze eerst geïnventariseerd moeten worden. Het instituut leende zich daarom uitstekend voor het ADA-project en was ook zelf tot medewerking bereid.

Conclusies en aanbevelingen op grond van het ADA-project biedt hoofdstuk 5, terwijl het laatste hoofdstuk tenslotte de ADA-aanpak beschrijft die uit dit project voor de toekomst als perspectiefrijk naar voren is gekomen.

2 Recente ontwikkelingen in de langetermijnbewaring

2.1 Het aandachtsgebied

Bij de oriëntatie voor dit project op ontwikkelingen in de digitale archivering is gekeken naar drie categorieën 'aangrenzende' instellingen: de wetenschappelijke data- en tekstarchieven, de bibliotheken en de overheidsarchieven. Het bewaren en ontsluiten van het cultureel erfgoed is voor elk van deze drie een belangrijke taak, soms zelfs de belangrijkste.

Door de snelle ontwikkelingen van de laatste jaren zijn veel instellingen gedwongen te gaan nadenken over de juiste vorm van bewaring en ontsluiting van het nu digitaal geworden cultureel erfgoed. In een aantal gevallen zijn pilotprojecten gaande en hier en daar is al met concrete preserveringsprogramma's begonnen. Helaas is er in de meeste landen tussen deze verschillende initiatieven weinig coordinatie, mede door de eerder genoemde verschillen in bewaarcontext.

Die verschillen kunnen echter niet wegnemen dat het bij de complexe digitale bewaringsproblematiek voor een belangrijk deel om dezelfde problemen gaat. Het duidelijkst komt dat tot uiting bij de verschillende bewaarstrategieën. Zo experimenteert de KB in het kader van het e-depot met langetermijnbewaring van elektronische documenten op basis van emulatie. De resultaten daarvan kunnen bruikbaar zijn voor veel andere erfgoedinstellingen. Ook andere aspecten als authenticiteit en integriteit van databestanden zijn in elke context van belang, al wordt er in de ene omgeving veel meer belang aan gehecht dan in de andere.

2.2 De internationale stand van zaken

2.2.1 Wetenschappelijke archieven

Al sinds de jaren zestig functioneren er archieven voor sociaalwetenschappelijke databestanden, voornamelijk in Europa en Noord-Amerika.[6] In de jaren tachtig zijn daar historische data-archieven en tekstarchieven bijgekomen. Hun belangrijkste taak is het bewaren en toegankelijk houden van bestanden, waarbij mo-

6 Voor een overzicht, zie <http://www.nsd.uib.no/cessda/europe.html>

gelijk hergebruik een belangrijk criterium is. De gehanteerde strategie is meestal conversie naar een software-onafhankelijk of gestandaardiseerd opslagformaat als ASCII of XML. Sociaal-wetenschappelijke data-archieven converteren meestal naar het SPSS-portable file format; tekstarchieven maken gebruik van de markup languages SGML of XML en historische data-archieven gebruiken ASCII en tegenwoordig op experimentele basis ook wel XML. Multimediale bestanden en op internet gepubliceerde databases vragen om nieuwe archiveringsoplossingen.

De data-archieven kennen een internationale standaard voor metadatasystemen, het Data Documentation Initiative van de internationale organisatie van data-archieven IASSIST. Dit DDI is geschikt voor verschillende soorten bestanden (database én tekstbestanden, multimediabestanden en websites) en kent lokale varianten. Het Nederlandse DDDI (Dutch DDI) leent zich voor het beschrijven van zowel sociaal-wetenschappelijke als historische databestanden.[7]

2.2.2 Wetenschappelijke bibliotheken

Vooral de nationale depotbibliotheken hebben zich noodgedwongen al sterk met de problematiek van de langdurige opslag van elektronische publicaties beziggehouden. Een aantal, zoals de Bibliothèque Nationale de France en sinds kort ook de KB, experimenteert met het archiveren van websites. In de wetenschappelijke wereld spelen *collaboratories* een rol, waarbij onderzoeksdata (zowel ruw als bewerkt), softwaretools en publicaties met een verschillende status op één website worden samengebracht.

In de wereld van de bibliotheken wordt vaak het OAIS-model aangetroffen. Dit Open Archival Information System is een referentiemodel voor alle relevante processen, zoals acquisitie, verwerking en ontsluiting van data. Er bestaan of ontstaan toepassingen voor specifieke bewaarcontexten, zoals voor depotbibliotheken het DSEP (Deposit System for Electronic Publications) dat in Nederland door de KB wordt gebruikt.[8] Ook de National Library of Australia gebruikt OAIS.

Vermeldenswaard is het CEDARS-project van de universiteiten van Leeds, Cambridge en Oxford om strategische, methodologische en praktische problemen op te lossen en handleidingen te maken voor wetenschappelijke digitale archivering. Ook dit project is echter sterk toegespitst op elektronische publicaties. Binnen CEDARS, dat overigens ook het OAIS-model als uitgangspunt heeft gekozen voor de langetermijnbewaring, is de data-archiveringsstrategie nog een belangrijk discussiepunt. Zo wordt met name over de emulatie-optie heel verschillend gedacht.[9]

---

7 Zie: <http://www.icpsr.umich.edu/DDI/index.html>

8 Voor het OAIS zie Dollar (1999) en voor het DSEP: <http://nedlib.kb.nl/>

9 Voor meer informatie over de CEDARS en CAMILEON projecten zie: <http://www.leeds.ac.uk/cedars/index.htm> en

2.2.3 Overheidsarchieven

Op het gebied van digitale archivering gebeurt bij overheidsarchieven wel het nodige, maar toch bestaat er een *major imbalance* tussen de verschillende archiefdiensten in Europa, zo bracht een onderzoek in opdracht van de Europese Commissie in 2001 aan het licht. In het noordwesten van Europa is men veel actiever op dit terrein dan elders. Veel werk wordt geïsoleerd verricht.[10] Op dit moment is er nog geen grootschalig digitaal overheidsarchief, zelfs niet in Noord-Amerika dat internationaal voorop loopt. De Amerikaanse National Archives and Records Administration (NARA) dat eveneens OASIS als basismodel gebruikt, heeft een aantal vaak grote projecten maar een grootschalige infrastructuur ontbreekt tot nu toe.[11]

Het archiveren van digitale archiefstukken (electronic records) stelt overigens andere eisen maar vraagt ook een andere organisatievorm dan die voor de andere twee categorieën. Dit verklaart ook de sterke toenadering van de laatste jaren tussen de archiefwereld en die van de documentaire informatievoorziening. Het streven is daarbij uiteindelijk tot een samenvoeging te komen van archief en DIV-afdeling.[12]

Interessant is de werkwijze van het National Digital Archive of Datasets (NDAD) in Londen, de digitale bestanden van de Britse overheid moet bewaren. De toegepaste methodiek is zeer praktisch en concentreert zich op het vaststellen van eenduidige protocollen en richtlijnen voor het beschrijven en overdragen van bestanden. De protocollen zijn zo ontworpen dat ze ook in gecompliceerde situaties gebruikt kunnen worden, waarbij verschillende partijen zeggenschap over de data hebben.[13] Net als in de ADA-opzet worden bestanden beschreven en overgedragen door een andere instantie dan die ze gevormd heeft.

In het archiefwezen is ISAD-G een internationale standaard voor metadata. Deze is sterk ontwikkeld op het terrein van het ontstaan en de vorming van archiefstukken met alle daarbij behorende bureaucratische processen, maar veel minder op het gebied van het documenteren van bestanden, met name ten aanzien van IT-specificaties.[14]

2.2.4 Grensoverschrijdende activiteiten

Er worden wel pogingen gedaan om de contextgebonden ontwikkelingen binnen deze drie categorieën instellingen beter te coördineren. Zo organiseert het DLM-Forum van de Europese Commissie een tweejaarlijks congres over langetermijnbewaring van digitale archiefstukken en tracht het standaardisatie tot stand te

10 Schürer (2001).
11 Zie voor een overzicht Thibodeau (2002).
12 Zie bijvoorbeeld Beagrie en Greenstein (1998).
13 Ashley (2002).
14 Uitgebreider hierover Shepherd en Smith (2000).

brengen door het uitgeven van *best practices*. En de Engelse Digital Preservation Coalition (DPC) functioneert nationaal als overlegorgaan voor de bibliotheek- en archiefwereld en de wetenschappelijke data-archieven op het terrein van digitale langetermijnbewaring.

In een aantal internationale projecten werken instellingen uit de verschillende categorieën samen. Op het gebruik van het OAIS is al gewezen. Het grootschalige InterPARES-project (International Research on Permanent Authentic Records in Electronic Systems) met de School of Library, Archival and Information Studies in Vancouver als hoofduitvoerder, tracht selectiemethoden en functionele eisen te formuleren, die authenticiteit van digitale documenten garanderen in de bibliotheek- en archiefsfeer.[15] Recent en veelbelovend is het Open Archives Initiative (OAI) dat zich richt op uitwisselbaarheidsstandaarden, met als primaire doelstelling om de toegang tot elektronische publicaties in elektronische bewaarplaatsen (*institutional repositories*) te verbreden. De gekozen werkwijze maakt het mogelijk dat databestanden verspreid zijn opgeslagen, terwijl uitwisselbaarheid wordt bereikt door een verplicht formaat voor metadata: de Dublin Core Metadata Element Set, gestructureerd als XML-document. Het OAI is in eerste instantie opgezet voor een gemakkelijke uitwisseling van e-prints, maar kan in principe tot alle soorten elektronische documenten uitgebreid worden. Daarbij wordt onderscheid gemaakt tussen *data providers* (die één of meer repositories onderhouden) en *service providers* (die de metadata van de data providers gebruiken en toegankelijk maken).[16]

2.3 De Nederlandse situatie

In Nederland is een aantal interessante ontwikkelingen gaande, maar deze hebben geen betrekking op onderzoeksbestanden. Zo acquireert het e-depot van de KB alleen digitale publicaties. Het e-archiving project van de Universiteitsbibliotheken van Delft, Utrecht en Maastricht ontwikkelde een interessante XML-containeroplossing voor met name elektronische publicaties[17], en het Archipol-project van de Rijksuniversiteit Groningen richtte zich op de websites van de politieke partijen – een voor Nederland uniek project voor webarchivering.[18] Het programma EDDA (Effectieve Digitale Duurzaamheid Amsterdam) van het Gemeentearchief Amsterdam kent in hoofdlijnen eenzelfde doel als het ADA-project, maar wordt anders uitgewerkt omdat het is gericht op het overbrengen van bestanden van de lopende administratie naar het statische archief.[19]

De Rijksoverheid heeft in oktober 2001 het *Testbed Digitale Bewaring* in het leven geroepen om 'de toegankelijkheid van betrouwbare overheidsinformatie in

15 Zie <http://www.interpares.org/>
16 zie < http://www.openarchives.org/ >
17 <http://www.library.tudelft.nl/ws/b/about_the_library/strategy/strategic_projects/earchiving/index.htm >
18 <http://www.archipol.nl/ >
19 <http://gemeentearchief.amsterdam.nl/concerndiensten/hulpmiddelen/edda/index.nl.html>

het digitale tijdperk te waarborgen, nu maar ook in de toekomst'. Het programma heeft een aantal publicaties opgeleverd waarin vooral softwareformaten zijn getest op duurzaamheid. Het heeft vooral betrekking op administratief-bestuurlijke bestanden en processen.[20]

Het door SURF gecoördineerde DARE-programma, waarin de Nederlandse universiteiten, de KB, de KNAW en NWO samenwerken, kwam in de laatste fase van het ADA-project op. Op basis van het hierboven beschreven OAI-model zijn in dit kader *institutional repositories* opgezet om wetenschappelijk onderzoeksmateriaal wereldwijd toegankelijk te maken, dus zoveel mogelijk binnen het publieke domein.[21] Binnen DARE-projecten zijn ook drie data-projecten uitgevoerd: op het terrein van de archeologie (e-depot Nederlandse Archeologie: eDNA), de onderwijskunde (Data Onderwijskundig Nederland Online Research: DONOR) en de hydrologie (Data Archiving River Environment Luxemburg: DareLux).

20 <http://www.digitaleduurzaamheid.nl/home.cfm>
21 DARE staat voor Digital Academic Repositories. Zie: <http://www.surf.nl/themas/index2.php?oid=18>

3 De case study 'Meertens Data'

3.1 Inleiding

Dit hoofdstuk bevat een verslag in hoofdlijnen van de werkzaamheden in de case study 'Meertens Data'. In dit verslag wordt de volgorde aangehouden zoals die bij de deponering van bestanden in het data-archief wordt doorlopen: inventarisatie, selectie, archivering inclusief beschrijving (het toekennen van metadata). Deze aanduidingen zijn weliswaar gangbaar binnen het data-archief, maar ze wijken af van de in de klassieke 'papieren' overheidsarchieven gebruikte terminologie.

Inventariseren betekent een overzicht maken van alle bestanden. Daarvoor is enige technische, algemene, inhoudelijke en organisatorische basisinformatie nodig. Zulke summiere documentatie geeft een eerste inzicht in de archiveringsmogelijkheden en maakt selectie in grote lijnen mogelijk. In de volgende fase, de selectie, wordt bepaald welke bestanden wel en niet gearchiveerd worden. Archiveren is dan het toegankelijk maken en houden van de bestanden: opslag maar ook beschrijven van de data. Dit laatste gebeurt door metadata toe te kennen met technische en meer contextuele informatie.

3.2 Inventarisatie

3.2.1 Overzicht van het materiaal en de data-infrastructuur

Om een goed beeld te kunnen krijgen van het digitaal erfgoed van het Meertens Instituut was inzicht nodig in de aanwezige data-infrastructuur – de bestanden, de media, de software en de hardware – en haar geschiedenis.

Begin jaren tachtig zette het Meertens Instituut de eerste stappen op het gebied van de automatisering. De toenmalige afdeling Dialectologie begon toen via een terminalverbinding zijn wetenschappelijke data te verwerken in samenwerking met het computercentrum van de Universiteit van Amsterdam. In lijn met het gebruik bij andere taal- en letterkunde-disciplines aan de Nederlandse universiteiten werd kort na de intrede van de *personal computer* overgestapt op het Macintosh-platform, dat er ook nu nog is. Vanaf omstreeks 1988 kreeg het instituut de beschikking over lokale afdelingsnetwerken, die geleidelijk zijn uitgebreid en aaneengebreid tot één AppleTalk netwerk. Eind 1995 werd een structurele upgrade van het netwerk uitgevoerd.

Tabel 3.1 Aantallen gegevensdragers naar soort

Drager	Aangeleverd	Uitgevallen	Geïnventariseerd
Diskettes 3½″	1497	36	1461
Diskettes 5½″	68	14	54
Diskettes totaal	1565	50	1515
Magnetische tapes	4	1	3
CD-ROMS	23		23
SyQuest back up-media	6		6
DAT medium	1	1	
Totaal	1599	52	1547

Een eerste overzicht van het grote aantal relevante bestanden bood het enige jaren daarvoor gemaakte inventarisatierapport *Gouden eieren*.[22] De bestanden die nog actief werden gebruikt bevonden zich op de server; die waarvoor dat niet (meer) gold op diverse losse media (tabel 3.1). Terwille van de overzichtelijkheid is er in overleg met het Meertens Instituut voor gekozen het onderzoek te concentreren op de laatstgenoemde groep bestanden. Deze groep, bij het Meertens Instituut gewoonlijk aangeduid als het materiaal uit 'de kast', is verspreid over een groot aantal uiteenlopende media waarvan het merendeel 3½″ diskettes (tabel 3.1).

Bij het Meertens Instituut werden deze data als afgesloten beschouwd, maar wel van belang geacht. Het ADA-project zou duidelijkheid verschaffen: welke bestanden zijn het waard om bewaard te worden en welke niet?

De variëteit aan bestanden bleek groot. Hoewel de bedoeling was geweest om alle bestandssoorten te verwerken, zijn de audiodata en de images uiteindelijk niet in het project betrokken. Bij de eerstgenoemde ging het vaak om min of meer commerciële producten (cd's met volksmuziek) of om digitaal gemaakte opnamen uit het eigen archief (spraak, vraaggesprekken en gezongen opnamen[23].) Voor de digitale archiefopnamen geldt dat de organisatie goed is toegerust om te reageren op calamiteiten, zoals selectief verslechterende deelverzamelingen. Deze kunnen snel worden geselecteerd om dan opnieuw te worden overgezet op een nieuwe drager.

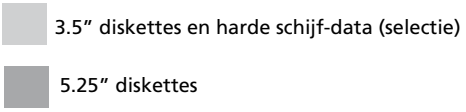
Daarnaast valt ook het aanbod aan beeldbestanden te verwaarlozen – niet verwonderlijk gezien de datering van het materiaal uit 'de kast'. Eind 2002 werd nog een aanvullende verzameling met jongere data in het ADA-project opgenomen: publieksgegevens van het project 'Brieven aan de Toekomst' uit 1998. Hieronder bevond zich wel een aantal beeldbestanden.

Het aanwezige materiaal bleek dus in grote lijnen overeen te komen met de verwachting, die voor een belangrijk deel gebaseerd was op het 'Gouden Eierenrapport'. Naast software in alle soorten en maten bestond het digitale materiaal voornamelijk uit tekstbestanden of data van gestructureerde aard (databases).

22 *Gouden eieren* (1997).

23 Deze laatste verzameld door Ate Doornbosch.

1980	20	16
1981	2	_
1983	_	1
1984	19	75
1985	_	16
1986	_	10
1987	24	109
1988	86	152
1989	175	223
1990	50	8
1991	81	35
1992	154	12
1993	147	2
1994	28	_
1995	244	2
1996	247	_
1997	128	_
1998	_	_



Figuur 3.1 Meertens Data: aantal bestanden naar jaar, waarin het jaar is afgeleid van de wijzigingsdatum.

3.2.2 Het inventarisatieproces

In de fase van de inventarisatie wordt een overzicht van alle bestanden gemaakt. De opgenomen informatie moet de basis vormen voor de later te maken keuzes bij de classificatie en de selectie. Er waren verschillende soorten gegevensdragers of media in het spel, te weten:
– Mac-data
– MS-Dos data
– Data op andere media of platforms
Elk van deze drie trajecten wordt hierna in hoofdlijnen besproken.

Mac-data
Van november 2000 tot april 2001 duurde bij het Meertens Instituut de inventarisatie van de oude digitale data op het Apple-Macintosh-platform. De NHDA-onderzoekers leverden slechts het ontwerp voor een datatabel, gebaseerd op het bij hun instituut gebruikte documentatieschema (DDDI) met beschrijvingselementen per bestand.

Het te inventariseren materiaal was verspreid over vele honderden 3½″ diskettes, die werden genummerd en vervolgens in de datatabel ingevoerd samen met gegevens over de inhoud. Volgens plan werd begonnen met de invoer van gegevens per bestand maar al vrij snel werd overgestapt op een hoger beschrijvingsniveau, de dataset. Dat kon inhouden dat een map (directory) met bij elkaar behorende bestanden als 'dataset' werd gedocumenteerd. Maar deze eenheid kon nog ruimer worden opgevat. Zo werden ook tot eenzelfde project behorende databestanden,

die op één diskette of op een reeks diskettes waren geplaatst, als één 'dataset' ingevoerd. Dezelfde werkwijze werd ook gehanteerd voor de meeste softwarepakketten.

De informatie die in een record is verzameld, kan zodoende betrekking hebben op één bestand maar ook op een verzameling van bestanden – een gebrek aan eenduidigheid dat in de latere fasen soms moeilijkheden bleek op te leveren bij de analyse en beoordeling van deze databank.

Eind 2001 waren er 1938 eenheden, met in totaal 7.406 Mb aan data, bij het Meertens Instituut verwerkt en gedocumenteerd. Later bleek dat dit uiteindelijk verreweg het grootste bestanddeel was van de door het Meertens Instituut aangeleverde data. Helaas is het, door de verder gevolgde werkwijze, niet mogelijk gebleken het uiteindelijk totaal aantal bestanden in deze inventarisatie te reconstrueren.

MS-Dos data: inventarisatie bij het NHDA

Van de overige media bleken alleen de MS-DOS geformatteerde floppy's verwerkbaar. In mei 2001 ontvingen we de eerste ca. zestig oude 5¼″ floppy-disks. In de eindfase van het project, eind 2002, werd ook nog een pakket 3½″ diskettes in behandeling genomen. In de voorselectie werden elf diskettes uitgefilterd vanwege de aard van hun gegevens.[24] Van de rest werd eerst bepaald of ze onderzoeksdata bevatten, om ze als waardevolle bestanden te kunnen indelen in de groep 'Meertens Data', dan wel in een restgroep met bestanden waarvan de relatie met het ADA-thema twijfelachtig leek. De definitieve selectie daarvan vond uiteraard later bij het Meertens Instituut plaats. Een handvol van deze floppies kon niet gelezen worden en is doorgespeeld naar het Computer Museum (zie III). Uiteindelijk bestond de subset in totaal uit 701 bestanden en 13,2 Mb aan data.

De later verwerkte dataset van ongeveer 45 3½″-diskettes bevatte de publieksreacties op het project 'Brieven aan de Toekomst'. Deze data zijn semi-automatisch geïnventariseerd. Totalen van deze subset: 439 bestanden, 9976 Kb.

Data op andere media of platforms: Computer Museum en Inventarisatie NHDA

Een aantal gegevensdragers behorend bij verouderde of in onbruik geraakte media die niet meer bij het Meertens Instituut of het NHDA konden worden gelezen, is bij het Computer Museum van de Universiteit van Amsterdam gebracht in de hoop dat ze daar weer toegankelijk gemaakt konden worden. Het ging om de volgende media:

24 Hieronder vallen onder meer versies van de NS-reisplanner, werkkopieën van het DOS PC-besturingssysteem. Ofschoon niet in lijn met het eerder omschreven uitgangspunt was het evident dat dergelijke minder relevante dataverzamelingen de procedure slechts zouden vertragen.

– *SyQuest removable harddisk cartridges* Eind 1995 heeft op het Meertens Instituut een vernieuwing van het computernetwerk plaatsgevonden. Het is hoogstwaarschijnlijk tegen deze achtergrond dat het gebruik van deze 'zip'-media als extra back-up faciliteit valt te verklaren. Analyse van de aanwezige directorystructuren levert de schatting op dat het om een kopie gaat van tien à twaalf harde schijven. In totaal bevatten deze zes back-up media 10.566 bestanden (695.120 Kb).

– *Mac 3½″ diskettes* Een klein deel van de 3½″ diskettes vertoonde schijffouten; deze zijn aan het Computer Museum doorgegeven. Twee konden alsnog worden gelezen. Totaal: 34 bestanden (1003 Kb).

– *Magneetbanden (mainframe computer SARA)* Van drie van de vier nog aanwezige magnetische tapes kon de inhoud gekopieerd worden. Eén exemplaar bleek leeg. De bitstream indeling[25] op de tapes zou op de PC de informatie slechts als een ongestructureerde brij weergeven. Daarom zijn de gegevens omgezet naar ASCII en in tabelvorm gestructureerd. Deze banden bleken drie databestanden te bevatten met in totaal 39.486 Kb aan gegevens.

– *Floppy-disks (5¼″)* Van deze vijf floppies bleken er twee leesbaar. De twee overige diskettes behoorden tot dezelfde set met een afwijkend platform: geformatteerd als Digital Rainbow CP/M.

In juli 2001 is het materiaal door het Computer Museum in leesbare vorm overgezet op CD-ROM, in Mac-formaat. In totaal zijn zo 10.603 bestanden extern gerecupereerd, met een gezamenlijke omvang van ca. 718 Mb.[26]

Bij de terugkeer van dit materiaal leidde extrapolatie van het grote aantal bestanden naar de ruim tien maal grotere reeds aanwezige set van Mac-data tot een schatting van in totaal meer dan honderdtien duizend aanwezige bestanden. Met het oog op dit grote aantal moesten vervolgstappen waar mogelijk worden geautomatiseerd. Catalogi van de digitale informatie werden dan ook zoveel mogelijk automatisch gemaakt, in de vorm van lijsten met bestanden of van een numerieke samenvatting met aantallen bytes, submappen en bestanden per directory. Deze uitvoer vormde de kern van een tabel met metadata. Met behulp van de hierbij ontwikkelde aanpak kon een snelle inventarisatie worden gerealiseerd.

Tabel 3.2 geeft de totalen van alle geïnventariseerde data, met de aantallen die na de eerste selectie overbleven. Clusters zijn groepen bij elkaar horende bestanden; meer hierover in paragraaf 3.2.3.

25 C.D.C. Display Code, 1600 bpi.

26 De set 5¼″ floppies niet inbegrepen.

Tabel 3.2 Resultaten werkproces Inventarisatie: fase I

Omschrijving / eenheid	Geïnventariseerd	Over na selectie
Data in Mb	8136	321
Clusters	1460	323

Van slechts 524 van de in totaal 1460 clusters zijn de bestanden geïnventariseerd. Het ging daarbij in totaal om 18480 bestanden, waarvan er na selectie 2979 overbleven. De overige 936 clusters zijn nooit verder geïnventariseerd.

Aan het einde van de inventarisatiefase was duidelijk geworden dat de selectie een iteratief proces is: om goed te kunnen selecteren moet vaak verdergaand worden geïnventariseerd, teneinde zekerheid te krijgen om welk bestand het gaat. Een verantwoorde selectie is afhankelijk van goede inventarisatiegegevens. Daarbij bleek een praktische procedure van groot belang. Daarin spelen drie processen een rol:
– clustering van data (paragraaf 3.2.3),
– classificatie (paragraaf 3.2.4),
– iteratieve bewerking.

3.2.3 Clustering van de data

Het onderscheiden van de afzonderlijke bestanden en hun begrenzingen bleek een lastig proces. Wat als bestand gezien kon worden, of als een groep bestanden, kon per medium verschillen. Op losse gegevensdragers waren de grenzen eenvoudig aan te geven: meestal is de floppy zelf de eenheid of dataset. Voor informatie op harde schijf ligt dit anders, en vooral bij grote gegevensverzamelingen zoals computer back-ups leverde de afbakening van de dataset een probleem op. Inventarisatie op twee niveaus, eerst op hoger 'dataset'-niveau en dan op bestandsniveau, was vereist.

Ten behoeve van het hogere niveau zijn de objecten geclusterd. Afzonderlijke onderdelen van een back-up, subdirectories met gelijkwaardige informatie, werden gehergroepeerd tot 'data clusters', die vervolgens bij het inventarisatieproces het eerste niveau van beschrijving vormden waarbij de oorspronkelijke padstructuur intact bleef. De clustering werd uitgevoerd door het Meertens Instituut zonder dat daarbij vaste criteria bestonden. Minimaal gold een drietal normen: de logische relatie van het object tot nabij liggende directories of submappen, een functionele toetsing om te beoordelen of de informatie tot dezelfde categorie behoorde (paragraaf 3.2.4), en de herkomst: de eigenaars of auteurs van de bestanden waren soms herkenbaar in de naamgeving.

Het resultaat was de zogenaamde clustertabel (onderdeel van de BIOM catalogus, zie bijlage C). Het toekennen van metadata en het waarderen en uitvoeren van andere bewerkingen kon nu worden toegepast op veel minder eenheden van

meestal grotere omvang. De informatie van de twaalf computer back ups is bijvoorbeeld in circa 200 data clusters opgedeeld en niet in 5000 directory-items.[27]

3.2.4 Classificatie van de data

Zowel in de inventarisatiefase als bij de selectie was het gebruik van een classificatie belangrijk. Onderbrengen van de gegevens in verschillende categorieën was nodig voor een snelle selectie. Daarbij is een eerder voorstel tot indeling, afkomstig uit het rapport *Digitaal Academisch Erfgoed* [28], uitgebreid en aangepast. Het classificatiesysteem dat zo ontstond is weliswaar toegesneden op dit project, maar kan goed dienst doen in vergelijkbare projecten.

Bij de classificatie in 'datagroepen' (zie tabel 4.3) is uitgegaan van een hoofdindeling in programmatuur (P), gecreëerde data (D) en centrale of systeem backups (SB).

Omdat het project primair gericht was op de onderzoeksdata, is hiervoor een systeem op maat gemaakt: de categorie D. Deze allerminst homogene groep is vervolgens verder onderscheiden in een aantal functionele klassen, ondermeer vanuit beheersoogpunt: in de eerste plaats de 'echte' onderzoeksdata, daarnaast een verzamelcategorie met van de eerste groep afgeleide producten en teksten. In een later stadium is een aparte klasse bron- of archiefdata onderscheiden: digitale informatie die voor toekomstig onderzoek benut kan worden. Hierbinnen viel tenslotte onderscheid te maken naar herkomst: interne en extern gecreëerde data.

De categorie IN bevat als enige zowel data als software. Voor ons doel was het niet nodig om de informatie omtrent beheer van het instituut nog verder uit te splitsen. Buiten het DX-materiaal zijn er op clusterniveau geen andere extern vervaardigde data gevonden; een enkele keer wel tekst maar dat leidde niet tot een afzonderlijk datacluster.

Ondanks alle aandacht voor (onderzoeks)data is de classificatie van de software zeker niet onderschat. Veel van de digitale erfenis had betrekking op de werking van randapparatuur of systeem- of ontwikkelsoftware. De inventarisatie zou hier kunnen volstaan met een beschrijving op het niveau van de cluster. Toch gold dit niet automatisch alle aangeleverde programmatuur. Voor het behouden van onderzoeksdata moet een maatwerkprogramma nu eenmaal anders behandeld worden dan systeemsoftware. Het bewaren van bijvoorbeeld een dBase-tabel zonder de bijbehorende maatwerkapplicatie (prg), kan gegevensverlies betekenen. Zeker als dit programma data verwerkt uit meerdere gekoppelde tabellen, zou ook

27 De nieuwe dataclusters kenden grote verschillen in omvang en in aantal onderdelen. Zo kon in een enkel cluster de halve directorystructuur van een harde schijf back up zijn opgenomen, omdat alle informatie in deze vertakking bij het Meertens Instituut als gelijkwaardig werd beschouwd (de ontwikkelomgeving bijvoorbeeld, of systeem-back ups). De kleinste eenheid daarentegen werd gerepresenteerd door een cluster bestaande uit slechts één file.
28 Mostert e.a. (1998) 11-13.

Tabel 3.3 Classificatie van soorten data

Aanduiding	Omschrijving
	P: Programmatuur (software)
PM	Maatwerkprogrammatuur: 'lokaal' door instituutsmedewerkers gecreëerd, of in opdracht geschreven. De relatie met data uit de groep DO is groot.
PX	Applicaties of kantoorsoftware. Commercieel geproduceerd en meestal op ruime schaal gedistribueerd. Van belang als de software-applicaties van oudere data zijn en indien incourant, met het oog op de conversie.
PS	Systeemsoftware. Commercieel en op grote schaal gedistribueerd, ten behoeve van de besturing van computersystemen (servers en PC's). Geen of beperkte relatie met onze aandachtsgroepen onderzoeksdata e.d.
PU	Verzamelgroep van utiliteitsprogrammatuur. Niet in alle gevallen zal de scheiding met PS duidelijk zijn. Geen relatie met onderzoeksdata e.d
	D: Digitale data, gecreëerd bij het Meertens Instituut (of elders)
DO	Onderzoeksbestanden: de wetenschappelijke 'output'. In het algemeen zijn het gestructureerde alfanumerieke gegevens, de ruwe uitkomst van het onderzoek. Vorm: databank, tabel, rekenblad.
DA	Archiefmateriaal: digitale informatie die de bron kan vormen voor onderzoekers. Heterogeen qua vorming, maar altijd betrokkenheid met het Meertens Instituut. In deze groep ging het om veel zelf of in opdracht gedigitaliseerde bronteksten (images en OCR-versies). Daarnaast documenten in het kader van het project van ingestuurde brieven. Ook transcripties van interviews.
DX	Extern vervaardigde, commercieel gedistribueerde data, vooral wetenschappelijk apparaat: bibliografieën, woordenboeken, soms ook wetenschappelijke bronteksten: bijv. 'Cetedoc' (Brepols).
DT	Teksten van wetenschappelijk medewerkers (ten behoeve van een proefschrift of andere publicatie). Ook uit databases afgeleide informatie (uitsnedes, export subsets).
DM/IN	Gegevens in relatie tot het (dagelijks) beheer van het instituut. Geen homogene groep; zal niet alleen data (teksten) bevatten maar ook lokaal en/of extern vervaardigde maat-applicaties (PX, PM).
DP	Persoonlijke mappen van medewerkers.
	SB: Systeem back-ups
SB	Directories met back ups van grote eenheden. Deze categorie bevat per definitie dus sterk heterogene groepen data, in het algemeen bestaande uit oude kopieën.

deze toepassing bewaard moeten blijven. Daarom is ook de programmatuur onderverdeeld. Bij het klasseren van een data cluster was het van belang om te weten met wat voor programmatuur (welke P-code) deze tot stand waren gekomen.

Van de in totaal ongeveer veertig diskettes met schijf- en of leesproblemen zijn er twintig in de databank opgenomen. Het zijn in alle gevallen 3½″ diskettes. De technische oorzaken van het meer en meer onleesbaar worden van data, de erosie noch de *data recovery* zijn bij dit project een substantieel aandachtspunt geweest, vooral vanwege het relatief weinig voorkomen daarvan.

Tabel 3.4 Totalen naar soort data

Soorten data: hoofdgroepen	Clusters	Bestanden *	Kb	% Kb
Onleesbaar	20		1925	0,02
Data	981	5881	942000	11,3
Programma's	405	10355	7168227	86,0
Systeem Backups	53	2250	219879	2,6
Totaal	1459	18480	8.332.031	100,0

* De aantallen in deze kolom geven de onvolledige gegevens uit de databank weer. De onvolledigheid is het gevolg van het feit dat de inventarisatie is gebaseerd op slechts 524 van de 1460 clusters, zie het vermelde bij tabel 3.2. Zie bijlage B voor een gespecificeerde versie van deze tabel.

3.3 Selectie

3.3.1 Criteria

De feitelijke selectie werd in twee fasen uitgevoerd. Eén voor het beoordelen per cluster, waarbij de eerder toegekende classificatie een nuttige rol speelde, en één op het bestandsniveau. De in deze tweede fase geselecteerde bestanden zijn uiteindelijk vrijwel allemaal gearchiveerd.

In het hele project ging het om het veiligstellen van bestanden met onderzoeksdata. Daarvoor zijn criteria nodig, die alleen maar door vakgenoten vastgesteld kunnen worden. Het Meertens Instituut heeft in dit project zelf de bewaarcriteria bepaald. Uit de eerste fase van de inventarisatie bleek de werkelijkheid al ingewikkelder te zijn dan het simpele onderscheid tussen wel of geen onderzoeksdata.

Het onderzoeksplan voor 2000-2005 van het Meertens Instituut[29] leverde 'urgentie en belang voor het wetenschappelijk onderzoek' als belangrijk criterium op. Het 'Gouden eieren'-rapport van hetzelfde instituut[30] noemt als voornaamste criterium het belang voor het lopende onderzoek binnen het Meertens Instituut en op de tweede het belang voor (samenwerking met) andere instellingen als KNAW-instituten of universiteiten. Daarnaast zijn er nog het innovatieve belang , het belang van conservering en tenslotte het maatschappelijk belang: behoud van en toegang tot cultureel erfgoed. Ook vanuit het NHDA zijn enkele criteria ingebracht zoals het belang voor onderzoek naar de langetermijnbewaring van digitale data.

3.3.2 Selectie fase 1: data clusters

Zoals eerder gezegd kon een deel van de eerste selectiefase automatisch uitgevoerd worden. Daarbij was selectie afhankelijk van de eerder toegekende classificatie. Zo vielen ondermeer clusters met de etiketten PS en PU af omdat deze hoogstwaar-

29 *Het oog op de toekomst, Onderzoeksplan 2000-2005* (1999), hoofdstuk 6.2(.2).
30 *Gouden eieren* (1997).

schijnlijk geen onderzoeksbestanden bevatten. Ook clusters met kantoorapplicaties, instituutsgerelateerde toepassingen en data en persoonlijke mappen van medewerkers werden uitgefilterd. Hetzelfde gold voor de SB-clusters, die kopieën van elders opgeslagen gegevens bevatten.

In de praktijk zijn niet alle clusters met onderzoeksdata (D) automatisch geselecteerd. Dat gebeurde bijvoorbeeld niet wanneer de informatie verouderd bleek of elders voorhanden.

De overwegingen voor de selectie zijn gedocumenteerd. Van het totaal van 1460 dataclusters werden er uiteindelijk 323 geselecteerd; deze bevatten circa 3000 bestanden.

3.3.3 Selectie fase 2: bestanden

Van de geselecteerde clusters werden vervolgens alle bestanden geïnventariseerd. Dat leverde een overzicht op met informatie over alle betrokken bestanden, ongeacht hun locatie of medium. Deze tweede fase van de selectie was dus gericht op deze bestanden. Dat selectieproces bestond uit drie niet altijd duidelijk te scheiden onderdelen of invalshoeken:

- technische uitfiltering op dubbele bestanden en oudere versies, waarna circa 2000 bestanden overbleven;
- inhoudelijke selectie door het Meertens Instituut, met als resultaat dat er in februari 2003 circa 700 bestanden geselecteerd waren voor bewaring;
- pragmatische selectie.

Deze laatste invalshoek hield in dat op grond van verschillende overwegingen[31] de samenstelling van de overblijvende verzameling nog aangepast kon worden. Enerzijds kon dat tot verdere selectie leiden, maar anderzijds ook tot deselectie. Per saldo is het aantal bestanden daardoor weer toegenomen.

Aanleidingen om bestanden alsnog uit te sluiten waren bijvoorbeeld:

- ze bleken alleen een lege structuur van een databank te bevatten;
- ze bevatten identieke informatie, opgeslagen in verschillende opmaak en met een andere bestandsnaam;
- het ging om font-bestanden;
- ontbrekende kennis van zaken.[32]

Na deze bewerking, waarvan ook het 'uitpakken' van enige zipfiles onderdeel was, zijn ongeveer 1300 bestanden geselecteerd voor archivering. Hiertoe dienden alle bestanden in non-ASCII-formaat eerst te worden geconverteerd. Deze conversie komt in de volgende paragraaf (3.4) aan de orde.

---

31 Afgezien van de eerder genoemde overweging van expertiseopbouw.

32 Het is in de loop van het project helaas niet mogelijk gebleken om data, gecreëerd met HyperCard en verweven met de programma-'stacks', hieruit los te weken. Vanuit het oogpunt van lange termijn bewaring is één van de denkbare oplossingen het converteren van de toepassing naar het programma MetaCard (met dank aan dr. E.H. Dooijes, Computer Museum UvA).

3.4 Archivering

3.4.1 Het werkproces in hoofdlijnen

De data-archivering was de derde en laatste fase van het werkproces. Deze fase zou moeten leiden tot een 'geoperationaliseerde infrastructuur' en 'gearchiveerde' databestanden: opgeslagen, gedocumenteerd en geschikt gemaakt voor raadpleging met behulp van de daartoe aangebrachte metadata.

De geselecteerde databestanden zijn met het oog op de langetermijnbewaring geconverteerd, met als standaard software-onafhankelijke ASCII-opslag[33], en gedocumenteerd met behulp van de database BIOM (Beheer en Informatie Oude Meertens-data). Deze bestaat uit twee gekoppelde hoofdtabellen: één op clusterniveau met contextinformatie over de gegevensdrager (het medium), en één op bestandsniveau met technische en contextuele metadata.

In het vervolg van deze paragraaf wordt nader ingegaan op de twee belangrijkste activiteiten van deze projectfase, de conversie en de documentatie.

3.4.2 Conversie: technische specificaties

Bij de aanvang van de conversie bedroeg het aantal geselecteerde bestanden om en nabij de 1300. Omdat voor een aantal geëxtraheerde tekstbestanden geen conversie nodig was, bleven 900 te bewerken items over.

De aard van de software, de programmatuur en het gebruikte platform, waarmee de bestanden zijn gecreëerd is van groot belang voor de organisatie van de conversie. Bij het beantwoorden van de vraag naar de gebruikte applicatiesoftware, was het platform de bepalende factor.

De data waren grotendeels afkomstig uit een Apple-Mac omgeving. Voor deze Mac-bestanden zijn daarbij in technisch opzicht twee begrippen van belang: de 'type'- en 'creator'-codes die de gebruikersinterface van het Mac besturingssysteem, de 'Finder', gebruikt om bestanden aan de juiste applicatie te koppelen. Samen worden deze onzichtbare codes ook wel de file signature genoemd.[34] Dankzij het feit dat deze signatures bij de inventarisatie zijn verzameld en als metadata in bestandentabel zijn opgenomen, hadden we in principe van tevoren inzicht in de aard van de Mac-bestanden.

---

33 Zie ook bijlage D (Tekst encoding).

34 Een bondige samenvatting hiervan (1): 'The Macintosh doesn't use the three-byte (or even more than three, like under Unix) extension concept to identify files, but signatures. Signatures are strings of eight bytes, four for the creator (the program which created the file) and four for the file type (text, picture, and so on). The correspondance between signatures and icons is managed by the Finder, for all programs which happened to exist on a volume, in the Desktop file (an hidden system file which is never shown by the Macintosh but exists on every disk).' (1) <http://www.macdisk.com/macsigen.php3 >. Voor meer informatie wordt verwezen naar de betreffende toolbox pagina op de apple.com site, 'Giving a Signature to Your Application and a Creator and a File Type to Your Documents',
< http://developer.apple.com/documentation/mac/Toolbox/Toolbox-447.html >

Tabel 3.5 Aantallen bestanden en spreiding naar platform en brontoepassing

Platform >> Toepassing	MS-DOS / Windows	Mac	Opmerking (*)
Wordperfect	200 *	6	meestal (95%) versie 5.1
FileMaker	0	143	-
WriteNow	0	169	-
Tekstbestanden	295	22	-
Word-documenten	19	4 *	Word 1.0 (WDBN WORD)
ClarisWorks 2.0	0	25 *	type CWDB, creator BOBO
Lotus 1-2-3	4 *	0	WK1 Lotus, release 2
Beeld-bestanden	10 *	0	jpg (9) en bmp
Ms Outlook- email	3 *	0	Outlook Express mail message (eml)
Totaal	531	369	-

Op een PC ontbreken deze file-attributen. Bij de groep PC-bestanden waren we dan ook aangewezen op de extensies. De vrijheid van naamgeving, zoals toegestaan door de verschillende softwareprogramma's, zorgde er echter voor dat de informatieve waarde hiervan beperkt was. Ruim 160 bestanden waren voorzien van een extensie-achtige toevoeging aan de naam, maar slechts bij 71 was de extensie te koppelen aan bekende software. Ook bij deze extensies, die schijnbaar aan een toepassing zijn gebonden, is echter niet alles wat het lijkt. Doc- en txt-bestanden bleken bijvoorbeeld met WordPerfect 5 te zijn gemaakt. Voor deze gevallen, voor de MS-DOS/Windows-bestanden zonder extensie en voor de bestandsnamen met een vrij toevoegsel van de auteur moest de applicatiesoftware proefondervindelijk worden vastgesteld. Dit leverde echter nauwelijks problemen op. Veelal was er snel een patroon te bespeuren en bleken clusters bestanden van dezelfde bron te bevatten.

Met al deze voorbehouden kon uiteindelijk een overzicht worden gegeven van de diversiteit in software van de originele data (Tabel 3.5).

3.4.3 Conversie: uitvoering

Alfanumerieke bestanden vormden dus, geheel volgens de verwachtingen, het leeuwendeel van de geselecteerde verzameling. Juist voor de archivering van deze groep bestanden was binnen het ADA-project in een oplossing voorzien. Dit gold veel minder voor de andere bestandsformaten. Voor de weinige beeldbestanden is ad hoc een oplossing gevonden.

Alfanumerieke bestanden

Zoals eerder uiteengezet vormde de strategie van migratie en conversie naar nieuwe systemen[35], bij voorkeur standaardformaten, in dit project het uitgangspunt voor het veiligstellen van de bestanden: het exporteren van bestanden in de originele software naar andere (nieuwere) software.

35 De terminologie wisselt soms in betekenis. Zie ook : Dollar (1999) en Bearman (1999).

In de praktijk hield dat de verwijdering in van alle door de software gegenereerde stuurcodes door middel van het opslaan van de informatie in standaard ASCII-formaat. De bestanden werden zo mogelijk in de eigen applicatie geopend en vervolgens opgeslagen of geëxporteerd als een tekstbestand. In een aantal gevallen kon een reeks (WordPerfect 5) bestanden worden omgezet als batch-opdracht met een conversieprogramma.[36]

Beeldbestanden
Buiten de groep alfanumerieke bestanden beschikten we over een zeer gering aantal beeldbestanden (9 jpg's en 1 bmp). De jpg-files zijn gedecomprimeerd en opgeslagen in uncompressed TIFF-format. Dit in overeenstemming met de huidige richtlijnen waarbij de jpeg-compressie als kwetsbaar wordt aangemerkt. Het bitmap-bestand is niet geconverteerd.

Ten behoeve van de feitelijke conversie heeft het NHDA een bescheiden computerlab ingericht. Het Meertens Instituut stelde een oudere, onder systeem 7 draaiende, Macintosh Performa 630 ter beschikking. Hierop draaiden de meeste aangetroffen Mac-toepassingen. Daarnaast hadden we voor de MS-DOS-bestanden de beschikking over een AT (een Hewlett Packard Vectra VL2) waarop naast Wordperfect 5.1 ook specifieke conversie-software was geïnstalleerd.

Tijdens deze fase van het project deden zich wat problemen voor door de relatief vrije naamgeving van de Mac-bestanden ten opzichte van het Windows-platform. Mogelijke problemen bij de geconverteerde doelbestanden zijn geneutraliseerd door het dichten van spaties en het vervangen van kritische tekens.

Hieronder een voorbeeld van een dergelijke ingreep: de oorspronkelijke Mac-bestandsnaam, links, en de gefatsoeneerde vorm van het conversiebestand rechts:

1. register (1959-1975)	1_register_1959_1975

3.4.4 Verdwijnende functionele opmaak

De gehanteerde conversie heeft nu en dan ongewenste gevolgen. Het beoogde resultaat, het verdwijnen van vrijwel alle stuurcodes, trof ook die bestanden waarbij markeringen van tekstblokken essentieel zijn voor de documentstructuur. Het ging hier om de uitgeschreven interviews met de vragen en opmerkingen van de ene partij gecursiveerd; de reacties van de geïnterviewde zijn zonder markering. In de geconverteerde versie is de leesbaarheid hierdoor duidelijk verminderd.

Voor deze groep bestanden – alleen tekstdocumenten – is de conclusie dat de hier uitgevoerde wijze van conversie niet kan plaatsvinden zonder de functionaliteit geweld aan te doen. De oplossing kan vermoedelijk alleen worden bereikt via een extra selectie- en bewerkingsslag, bijvoorbeeld een *zoek & vervang*-ingreep in de relevante teksten. Stuurcodes worden dan vervangen door eenvoudige markeringen, al dan niet gebaseerd op HTML (en , et cetera).

36 Software Bridge v.5.

3.4.5 Documentatie van de gegevens

Zonder documentatie is het niet mogelijk de data toegankelijk te maken. In het beginstadium van het project was er, in het kader van de inventarisatie, geen scheiding tussen de invoer van basisgegevens en aanvullende verrijking. Voor de data afkomstig van de harde schijf-back ups heeft dit laatste pas na de selectie plaats gevonden.

Speciaal voor dit documentatieproces is de eerder genoemde BIOM-database gemaakt (paragraaf 3.4.1) in de vorm van een Access-databank, bestaande uit twee gekoppelde tabellen waarin de metadata met betrekking tot respectievelijk de clusters en de bestanden zijn opgenomen. Details over de BIOM-catalogus geeft bijlage C.

Het Meertens Instituut zal zelf verdere verrijking van inhoudelijke aard uitvoeren, met deze catalogus als uitgangspunt. Het gaat om de in de bestanden of bestandsnamen besloten informatie over plaats en periode die via een geografisch zoekveld, op meerdere niveaus getrapt, en een periode-ingang toegankelijk worden gemaakt.

Naast de documentaire rol, die een onderdeel vormt van de archivering, kreeg de databank in de loop van het project een steeds belangrijker beheersfunctie. Zo vormde de in de clustertabel opgeslagen informatie in eerste instantie de basis voor de selectie. Ook het uitfilteren van identieke bestanden en verdubbelingen met verschillende datum kon hiermee eenvoudig worden gerealiseerd. Verder was BIOM een bron voor het maken van de kencijfers, die deels in dit verslag zijn opgenomen.

3.5 Conclusies

Naast concrete resultaten (900 gearchiveerde databestanden[37] en een databank met metadata) heeft dit project ook veel ervaring opgeleverd met de toegepaste werkwijze. Op een aantal punten bleek de werkelijkheid anders uit te pakken dan van te voren was gedacht. Dit noopte tot aanpassingen in de oorspronkelijke projectopzet. Het grote aantal bestanden was zo'n punt. De beheersing van de omvang van het project werd al vrij snel belangrijk en leidde als vanzelf tot een meer iteratieve aanpak. Ook de beschrijving van de data werd op een beperktere schaal aangepakt dan oorspronkelijk voorzien. De rol van de opdrachtgever was daarnaast bij vele beslispunten doorslaggevend. Dat speelde niet alleen bij inhoudelijke beslissingen, maar ook bij de ontsluiting en ter beschikkingstelling van de data. Een belangrijke constatering is verder dat dit project een vrijwel uitsluitend retrospectief karakter heeft gekregen. Dat was niet de bedoeling, maar het reconstrueren van de oudere bestanden bleek al ruim voldoende voor één project.

Een aantal van de belangrijkste bevindingen uit de praktijk van het project volgt hieronder puntsgewijs.

37 Plus ruim 350 geëxtraheerde tekstbestanden, waar verder niets aan hoefde te gebeuren.

1. De hoeveelheid data De inventarisatie en de selectie kregen door de grote hoeveelheid data een sterk iteratief karakter. Er werd eerst op het niveau van de dataclusters geselecteerd en daarna op dat van de databestanden. Zo kon 96% van de data al in de eerste fase worden uitgeselecteerd. Het aandeel van de uiteindelijk geselecteerde data bedraagt ongeveer 1%. Het zal duidelijk zijn dat een selectie op basis van goed ingedeelde clusters in de eerste fase veel onnodig werk later in het project kan voorkomen. De constatering dat inventarisatie en selectie sterk iteratief verliepen, heeft geleid tot een aanpassing van de ADA-aanpak (zie hoofdstuk 6). De juiste volgorde van een ADA-project moet zijn: eerst selectie op projectniveau, vervolgens op clusterniveau en vervolgens op bestandsniveau. Onnodig heen-en-weer springen tussen deze niveaus moet zoveel mogelijk voorkomen worden, al zal dat in de praktijk nooit helemaal lukken.

2. Inzicht in de data-infrastructuur en data-collecties De aangetroffen data waren naar vorm, medium en platform, en tot op zekere hoogte op afdelingsniveau verschillend. Bij het Meertens Instituut zelf kon het inzicht in wat er eigenlijk werd aangetroffen variëren. Dat kwam onder meer door de voorafgaande veranderingen in het platform, het operating system en de software, maar ook door omstandigheden als de aanschaf van computers zonder diskettestation of de beëindigde relatie met externe dienstverleners.

3. De technische staat van de data De technische staat van de data was over het algemeen goed. Het digitale materiaal was leesbaar was of kon zonder grote problemen leesbaar worden gemaakt. Een verwaarloosbaar percentage van de diskettes bevatte technische fouten. Er bleken op dit punt geen ernstige problemen, bijvoorbeeld als gevolg van verkeerde opslag of calamiteiten.

4. De benodigde inzet van de opdrachtgever Het contact met de opdrachtgever is onontbeerlijk geweest. Dat gold voor de selectie, het samenvoegen van mappen tot clusters, het inhoudelijk verrijken van deze data clusters onder andere met instituutsinformatie.

Het contact verliep bovendien bijzonder goed. Het project heeft veel profijt gehad van de nog steeds bij het instituut bestaande kennis over de organisatie en haar eigen verleden. Betwijfeld moet echter worden of de hier bestaande continuïteit nog als regel kan worden beschouwd in wetenschappelijk Nederland. Gelet op de veranderingen bij de universiteiten (schaalvergroting, samenvoeging en opheffing van onderzoekseenheden) in de afgelopen decennia, wordt de kans steeds kleiner dat de verantwoordelijke afdeling documentaire informatievoorziening nog in staat zal zijn inhoudelijke hulp te bieden bij de verrijking. Op grond van de in dit geval positieve ervaringen is de conclusie dat een zekere inzet van de opdrachtgever minimaal vereist is. Het gaat daarbij om inhoudelijke beoordeling van de databestanden en kennis van de huidige en de historische data-infrastructuur.

5. Wensen van de opdrachtgever De betrokkenheid van de opdrachtgever kan in intensiteit variëren, afhankelijk van wat deze partij uiteindelijk wil. In het kader van dit project had het Meertens Instituut tevoren geen specifieke wensen geuit. Het wilde primair inzicht hebben in de data, maar was daarbij niet direct in staat om selectiecriteria op te geven. Daarnaast is ook de fase van de ontsluiting der data, in het bijzonder mogelijke publicatie daarvan op Internet, in het kader van dit project verder niet uitgewerkt. Dit alles heeft er toe geleid dat de nieuw ontwikkelde ADA-aanpak een sterk modulair karakter kent.

6. Planning De eerdere projectfasen vergden veel meer tijd dan oorspronkelijk voorzien, de latere fase van de conversie vroeg juist veel minder tijd. Er was met name een behoorlijke discrepantie tussen plan en uitvoering bij de verwerking van de ruim 8 Gb aan data. Getracht is ook de kosten hiervan te becijferen.

Vanuit het oogpunt van planning moet ook rekening gehouden worden met de communicatie tussen opdrachtgever en uitvoerder, waarvan de intensiteit vanzelfsprekend grotendeels wordt bepaald door de overeengekomen mate van participatie van de opdrachtgever. In dit project is een deel van de tijd 'opgegaan' aan interne communicatie bij de opdrachtgevende instelling; het navragen van informatie bij betrokkenen ten behoeve van de documentatie. Daarnaast bleek de communicatie tussen uitvoerder en een externe dienstverlener (Computer Museum) van groot belang. Ook daarbij bleek een goed contact en ook controle nodig te zijn.

7. Technische punten Op enkele technische punten wordt in bijlage D ingegaan.

4 Het marktonderzoek

4.1 Inleiding

Het marktonderzoek, om te kunnen vaststellen of de wetenschappelijke wereld geinteresseerd is in digitale archiveringsdiensten, heeft zich geconcentreerd op instituten in de humaniora en sociale wetenschappen. Dat is het werkterrein van het NHDA[38] en zijn erfopvolger DANS. Hoewel daar niet expliciet naar is gezocht, is toch enige informatie over andere disciplines aan het licht gekomen. Daardoor is het mogelijk algemene conclusies te trekken over de situatie rond de Nederlandse onderzoeksdata.

4.2 Opzet

Gekozen is voor een aanpak op basis van diepte-interviews met vertegenwoordigers van een aantal instituten, gespreid naar soort onderzoek en discipline. *Onderzoeksinstellingen* hebben vaak zelf gegevensbestanden aangelegd, terwijl andere instellingen in de eerste plaats *documentatiecentra* zijn, die vaak grote collecties gegevens (documenten, teksten) beheren en uitgeven en daarnaast ook nog onderzoek verrichten. Beide categorieën produceren data en stellen deze meestal via een website of anderszins ter beschikking. Er is gesproken met zowel universiteiten als instituten.

Gesprekken zijn gevoerd met vertegenwoordigers van:

- het Sociaal Historisch Centrum Limburg in Maastricht,
- de Rijksdienst voor Kunsthistorische Documentatie (RKD) in Den Haag,
- het Instituut voor Nederlandse Geschiedenis (ING) in Den Haag,
- het Instituut voor Nederlandse Lexicologie (INL) in Leiden,
- het KITLV (Koninklijk Instituut voor Taal-, Land en Volkenkunde) in Leiden,
- het NIDI (Nederlands Interdisciplinair Demografisch Instituut) in Den Haag,
- de Fryske Akademy in Leeuwarden,
- de Theologische Universiteit in Kampen,

38 Dit is in afwijking van de oorspronkelijke opzet, waarbij een marktonderzoek in alle disciplines was voorzien. De ten opzichte van de beginfase van het ADA-project veranderde strategische omgeving van het NIWI heeft tot deze verandering genoopt. Ook DANS richt zich op dit moment uitsluitend op de alfa- en gammawetenschappen.

– WODC (Wetenschappelijk Onderzoeks- en Documentatiecentrum van het Ministerie van Justitie),
– de Universiteit van Amsterdam (archiveringsproject).
Ook met het CBS (Centraal Bureau voor de Statistiek) is contact geweest.

In de meeste gevallen werd gesproken met de verantwoordelijke voor het automatiseringsbeleid. Dat leverde inzicht op in de manier waarop digitale bestanden gecreëerd en gebruikt worden voor wetenschappelijk onderzoek, en in de geschiedenis van de automatisering van het instituut de afgelopen jaren.

De gesprekspartners werden van tevoren geïnformeerd over de opzet van het ADA-project en in het bijzonder over de mogelijkheden van digitale archiveringsdiensten. In het gesprek kwam een aantal vragen aan de orde. De eerste groep vragen was er op gericht te kunnen vaststellen of een instelling zelf inzicht heeft in de door haar of onder haar dak vervaardigde bestanden. Daaraan gekoppeld was de vraag of men zelf in de praktijk al tegen problemen met de leesbaarheid en bruikbaarheid van oudere bestanden was opgelopen.

De tweede groep van vragen was erop gericht, te weten te komen in hoeverre de instelling zich bewust was van de digitale duurzaamheidsproblematiek, daar zelf enig beleid voor heeft ontwikkeld dan wel van plan was te gaan ontwikkelen. De derde groep van vragen ging in op de digitale archiveringsdiensten en de in het ADA-project voorgestelde werkwijze.

4.3 Vragen over de inventarisatie van de bestanden

Het antwoord op de vragen naar het eigen inzicht in de aanwezige digitale bestanden was zeer divers. Sommige hadden, volgens eigen zeggen, hun bestanden goed op orde. Dit deed zich vooral bij die instituten voor waar men als belangrijkste doelstelling het uitgeven van digitale teksten of databestanden heeft (INL, ING). Wel wordt een duidelijk onderscheid gemaakt tussen eindbestanden en werkbestanden. De stellige indruk bij deze instituten is dat men zich de waarde van deze bestanden ten zeerste bewust is en zeker voor de technische back-up goede veiligheidsmaatregelen heeft genomen. Dat geldt zowel voor de eigenlijke data (of tekst-) bestanden als de daarbij behorende metadata. Daarnaast is vaak uitdrukkelijk voor zeer gangbare systemen gekozen (software en hardware).

Bij andere instituten was een minder duidelijk of gemengd beeld: in sommige afdelingen en bij sommige projecten wel, bij andere had men geen overzicht. Daar kwam een factor bij: niet in alle gevallen lijkt het gemakkelijk een onderscheid te maken tussen bestanden van het instituut en van individuele medewerkers, verbonden aan het instituut. Bedrijfsculturen kunnen op dit punt enorm uiteenlopen.

4.4 Besef van de duurzaamheidsproblematiek

Ook bij de vragen naar het besef van de problematiek van duurzaamheid was het beeld gevarieerd. In de hiervoor genoemde groep instituten die hun bestanden be-

ter op orde hadden kon wel besef van de bewaarproblematiek aangetroffen worden, gecombineerd met ongerustheid over de fysieke duurzaamheid van gegevensdragers (CD-ROMS, optische disks), de afhankelijkheid van *proprietary software*. Een paar keer werd het ontbreken van enig centraal beleid op dit punt genoemd.

Voor kleinere instituten blijkt de situatie anders dan voor grote organisaties, zoals universiteiten en bijvoorbeeld het Centraal Bureau van Statistiek. Bij kleinere instituten hebben systeembeheerders of vergelijkbare functionarissen nog enig overzicht. Zelfs daar wordt echter niet alles overzien, zeker niet de bestanden van de individuele medewerkers.

Er is zeker enig streven naar duurzaamheid, doordat men bijvoorbeeld *backward compatibility* toepast en een goed back-up systeem verzorgt. Ook zeiden sommige gesprekspartners de ontwikkelingen op dit terrein te volgen, speciaal wat er in Nederland gebeurt en met name bij het e-depot van de KB of het Programmabureau Digitale Duurzaamheid van de Rijksoverheid.

Bij een aantal instituten, vooral in de alfa-sfeer, is begrijpelijkerwijze de aandacht bijna volledig gericht op digitalisering van teksten of beeld. Dientengevolge wordt de bewaarproblematiek sterk onderbelicht, om niet te zeggen onderschat.

4.5 Houding tegenover de ADA-aanpak

Het werd duidelijk dat bij de meeste instituten het besef aanwezig is dat men zelf niet aan beleid voor langetermijnbewaring of de uitvoering daarvan toekomt. In hoeverre is er nu behoefte aan ADA-dienstverlening? In ieder geval bestaat er ongerustheid, en daarmee een vraag naar voorlichting, documentatie en richtlijnen op dit gebied. Daarnaast constateerden de meer betrokken respondenten twee problemen. Ten eerste kunnen landelijke initiatieven (nog) niet altijd gemakkelijk naar de praktijk van een instituut vertaald worden. In de tweede plaats wordt de uitkomst van het internationale debat over de beste archiveringsstrategie als onzeker beschouwd. Definitieve keuzes op dit terrein worden daarom op dit moment uitgesteld, dan wel voor zich uit geschoven. De keuze tussen emulatie of conversie wordt als te moeilijk gezien: er zijn nog geen overtuigende voorbeelden uit de praktijk.

De noodzaak van beleid op dit terrein werd over het algemeen wel ingezien. Desgevraagd konden vrij gemakkelijk allerlei *dark digital archives* genoemd worden: collecties van bestanden, waarvan soms zelfs betwijfeld werd of er nog wel iemand verantwoordelijk voor was. Ook hier is de schaalgrootte belangrijk: bij een universiteit zal zoiets zich eerder voordoen dan in een klein instituut. Daarnaast spelen reorganisaties een belangrijke rol. Daarbij verdwijnen nogal eens collecties geheel uit het zicht. Ook persoonlijke elementen als de aanwezigheid van systeembeheerders of documentalisten kunnen een factor zijn.

Een aantal respondenten zag inventarisatie en selectie als nuttige activiteiten, uiteraard vooral bij de instellingen die geen goed beeld hadden van hun bestanden én van het belang van inspanningen om deze te gaan bewaren. De vraag is echter

wel in hoeverre men dit een hoge prioriteit geeft. Zonder extra investeringen in tijd en geld zullen de bedoelde behoudswerkzaamheden waarschijnlijk niet van de grond komen.

4.5 Conclusies

Vast staat dat voor het onderwerp langetermijnbewaring grote belangstelling is en dat er behoefte is aan expertise op het terrein van de digitale archivering. Geconcludeerd mag worden dat de bewustwording zeker is toegenomen.[39] Een duidelijk beleid is er echter meestal nog niet, laat staan dat dit wordt uitgevoerd. Bij sommige instituten bestaat echter wel reeds een duidelijk beleid voor langetermijnbewaring, al of niet gekoppeld aan beschikbaarstelling via een website. Of dat beleid in alle opzichten adequaat is, is wel de vraag. Sommige gesprekspartners twijfelden hier zelf over. Bij een groot aantal instituten staat de wens van meer en snellere digitalisering en beschikbaarstelling van onderzoeksmateriaal voorop en daardoor worden de archiveringsproblemen onderbelicht en onderschat.

Op de concrete vraag in hoeverre er in wetenschappelijk Nederland behoefte bestaat aan de ADA-dienstverlening is geen eenduidig antwoord te geven. Enerzijds lijken de instellingen niet onmiddellijk uit zichzelf tot activiteiten over te gaan, of het nu archivering van oude bestanden betreft (zoals in dit ADA-project) of van het huidige materiaal. Anderzijds lijkt er wel degelijk een voedingsbodem aanwezig te zijn, zeker indien wanneer archivering gestimuleerd wordt door externe factoren zoals beleid van de kant van de universiteiten, landelijke onderzoeksorganisaties (NWO, KNAW) of de Rijksarchiefinspectie.

In ieder geval bestaat er een vraag naar voorlichting, documentatie en richtlijnen. Zeker inventarisatie en mogelijk daarop volgende selectie van bestanden (de eerste fasen van de ADA-aanpak) wordt door een aantal instellingen als gewenst gezien, ook met de financiële consequenties in beeld.

De indruk uit de contacten met de universiteiten is dat daar wel degelijk behoefte is aan expertise op het terrein van het ADA-project. Enkele gesprekspartners uitten ook de behoefte aan een landelijke instelling waar onderzoeksbestanden kunnen worden bewaard.[40] Op zijn minst lijkt een landelijke registratie van databestanden of althans de grotere collecties daarvan gewenst. Dat geldt echter ook ten aanzien van de categorie 'rondzwervende' bestanden van individuele medewerkers.

---

39 In vergelijking met enige jaren geleden, zie Mostert e.a. (1998).

40 Naast de hierboven genoemde data-archieven voor de sociale wetenschappen en geschiedenis lijken nu ook de neerlandici zich van de bewaarproblematiek bewust. Dat kan althans afgeleid worden uit de plannen zoals aangegeven in het in opdracht van de Nederlandse Taalunie uitgevoerde onderzoek *Blauwdruk voor onderhoud, beheer en distributie van door de overheid gefinancierde digitale materialen* (2002)

5 De haalbaarheid van digitale archiveringsdiensten

5.1 Conclusies van het ADA-project

In het ADA-project stond de vraag centraal naar de haalbaarheid van retrospectieve digitale archiveringsdiensten voor de wetenschappelijke wereld. Deze vraag wordt vanuit twee perspectieven beantwoord: vanuit het aanbod en vanuit de vraag.

Van de kant van het NHDA en zijn erfopvolger DANS – de aanbodzijde – is het antwoord bevestigend: de ervaring met het pilot-project heeft het mogelijk gemaakt dat deze vorm van dienstverlening aan wetenschappelijk Nederland aangeboden kan worden.[41] Duidelijk is geworden dat dit een dienstverlening op maat zal moeten zijn die altijd projectmatig opgezet moet worden. De situaties, mogelijkheden en wensen kunnen sterk uiteenlopen, wat ook voor een flexibele en modulaire – kortom projectmatige – aanpak pleit. In de beschrijving van de ADA-aanpak in hoofdstuk 6 worden de mogelijke vormen van deze dienstverlening verder uitgewerkt. De aanpak is voor alle wetenschappen geschikt en zou overigens ook buiten de wetenschappelijke sector aangewend kunnen worden.

Aan de vraagzijde bestaat aan advisering, voorlichting, kortom expertise op het terrein van langetermijnbewaring, zeker behoefte. Dat geldt ook voor het inventariseren van databestanden. Verdergaande activiteiten als selectie, archivering en ontsluiting hebben veel instituten zelf nog niet duidelijk in beeld. Uiteraard speelt bij alle ADA-activiteiten het financiële aspect een belangrijke rol, zeker wanneer het om grotere, arbeidsintensieve projecten gaat. Ook de grootte van de instelling is uiteraard een factor. De opdrachtgever kan echter door de modulaire aanpak en het inzetten van eigen menskracht een grote eigen inbreng hebben. Essentieel is dat het maatwerk blijft zodat de kosten van fase tot fase vastgesteld en beheerst kunnen worden.

41 Na de afsluiting van dit ADA-project is dit ook al daadwerkelijk gebeurd in het project e-depot Nederlandse Archeologie <http://www.edna.leidenuniv.nl/>, waarin archeologische bestanden in repositories zijn opgeslagen. In dit project is de ADA-aanpak gevolgd.

Het marktonderzoek heeft ook de urgentie van het probleem bevestigd. Het risico dat onderzoeksbestanden verdwijnen wordt snel groter door het verdwijnen van de kennis over de te archiveren data. Vooral het documenteren en selecteren van data, die hebben toebehoord aan inmiddels verdwenen vakgroepen of instituten kan veel problemen gaan opleveren. Het probleem lijkt eerder op het vlak van verdwijnende kennis te liggen dan op dat van de leesbaarheid van oudere media. In het proefproject bleek deze leesbaarheid nauwelijks een probleem, met uitzondering van een beperkt aantal diskettes met schijffouten. Bij digitale archivering tijdens of direct na de creatie van bestanden zal dit soort problemen zich veel minder voordoen. Daarom is deze actieve vorm van archivering, ongebruikelijk in de papieren archiefwereld, zo noodzakelijk om langetermijnbewaring van digitale bestanden te bewerkstelligen.

Nevendoel van het project was dat de 'resultaten (...) een bijdrage kunnen leveren aan de oplossing van de problematiek van het bewaren van digitale wetenschappelijke bestanden voor de lange termijn of, op zijn minst, aan mogelijk beleid in Nederland op dit terrein'. In het ADA-project is ervaring opgedaan met het samenspel van factoren, die een rol spelen bij het digitaal archiveren van databestanden. Deze zijn van technische, documentaire, maar vooral van organisatorische en beleidsmatige aard. Daarnaast is uit het marktonderzoek heel duidelijk de noodzaak gebleken van een (centrale) instelling in Nederland, die zich bezighoudt met advisering en mogelijk ook opslag van wetenschappelijke databestanden.

5.2 Aanbevelingen voor de ADA-aanpak

Op grond van de in het pilot-project ('Meertens data') opgedane ervaringen kon een aantal kritische aandachtspunten geformuleerd worden, die op hun beurt weer hebben bijgedragen aan de formulering van de ADA-aanpak voor toekomstige digitale archiveringsdiensten zoals in het volgende hoofdstuk weergegeven. Deze punten volgen hier in het kort:

1. Noodzakelijk is een goede schatting vooraf van de grootte van het project, vooral gebaseerd op het aantal databestanden en de variëteit daarin, inzicht in de data-infrastructuur en de automatiseringsgeschiedenis van het opdrachtgevende instituut.
2. De communicatie met het opdrachtgevende instituut en eventuele andere instellingen waarvan assistentie vereist is moet gedurende het hele project een belangrijk aandachtspunt blijven..
3. Een selectie in fasen kan latere overbelasting van de infrastructuur voorkomen.
4. Een modulaire aanpak is sterk aan te bevelen, vooral door de keuzemogelijkheden die worden geboden aan de opdrachtgevende instantie.
5. Participatie van de opdrachtgever is bij elke module noodzakelijk, maar de intensiteit daarvan kan sterk verschillen. Duidelijke afspraken met de opdrachtgever hierover zijn essentieel.

6. De medewerking van het opdrachtgevende instituut zelf is onontbeerlijk, speciaal ten aanzien van de inhoudelijke kennis van de databestanden en de context waarin deze zijn ontstaan. Zonder de inschakeling van vakspecialisten is inhoudelijke selectie niet mogelijk.
7. Inventarisatie en selectie kunnen een sterk iteratief karakter hebben. De selectie dient daarom zoveel mogelijk in lagen (project – cluster – bestand) uitgevoerd te worden
8. De ADA-aanpak moet generiek zijn en niet domeingebonden, en kan in principe zowel binnen als buiten de wetenschappelijke wereld toegepast worden.
9. De ADA-aanpak moet primair gericht zijn op digitale bestanden en niet speciaal op elektronische tijdschriften of boeken.

6 De ADA-aanpak voor digitale archiveringsdiensten

6.1 Inleiding

De ervaringen en conclusies van het ADA-project hebben de basis gevormd voor het formuleren van een meer algemene, toekomstgerichte aanpak voor digitale archiveringsdiensten. Die zogenaamde ADA-aanpak is echter niet alleen gebaseerd op de ervaringen in het pilot-project, maar ook op en de standaardwerkwijzen van het NHDA en op de werkwijzen zoals die worden aanbevolen in diverse handboeken, artikelen over *best practices* en *white papers* voor digitale archivering.[42] Tevens wordt waar mogelijk aangesloten bij de aanpak in de traditionele archiefleer en worden standaard archieftermen gebruikt voor zover van toepassing.[43]

De ADA-methodiek is gericht op retrospectieve archivering: het achteraf archiveren van digitale informatie en is minder geschikt om lopende projecten te archiveren. Het verdient aanbeveling om naast de uitvoering van een ADA-project gericht op de historische dataproductie, ook maatregelen te nemen voor prospectieve archivering.

De ADA-aanpak is in principe stapsgewijs en hiërarchisch van opzet, waarbij een top-down benadering wordt gevolgd. De methodiek is ook verwant aan de watervalmodellen uit de systeemontwikkeling, zoals System Development Methodology. Iedere te zetten stap in het proces is afhankelijk van informatie die in de vorige stap is verzameld en de keuzes die daarbij zijn gemaakt. Het is slechts beperkt mogelijk om terug te keren op eenmaal gemaakte keuzes. Het veranderen daarvan of het wijzigen van eerder gestelde prioriteiten zal consequenties hebben voor de tijd en kosten van het project.

De ADA-aanpak houdt in dat de selectie, beschrijving en documentatie van het digitale archiefmateriaal in een aantal iteraties plaatsvindt, waarbij gewerkt wordt van het algemene, globale niveau naar het specifieke, meer gedetailleerde niveau. Bij de uitvoering van een ADA-project kan men besluiten om de archivering te be-

42 Zie: Sheppard en Yeo (2003), Cox (2001), Jones en Beagrie (2001), Dollar (1992 en 1999) en Lazinger (2001).
43 Zie: Den Teuling (2003).

perken tot enkele niveaus. Uiteraard zijn de diepere niveaus van selectie, beschrijving en documentatie tijdrovender en dus kostbaarder. De stapsgewijze aanpak biedt echter een handvat voor verantwoorde selectie. De opzet van het werkproces is juist hiërarchisch om de selectie zo efficiënt mogelijk te doen plaatsvinden en om te voorkomen dat overbodig werk wordt gedaan. De zeven onderscheiden fasen van de voorgestelde ADA-aanpak zijn:

Fase 1: Omgevingsbeschrijving
Fase 2: Materiaalafbakening
Fase 3: Selectie en inventarisatie van projecten
Fase 4: Mediumconversie en documentatie op het niveau van dataclusters
Fase 5: Documentatie op bestandsniveau en conversie naar standaardformaat
Fase 6: Documentatie op gegevensniveau
Fase 7: Bewaring en toegankelijkheid

De ADA-methode is gericht op de archivering van (wetenschappelijke) dataproductie. Daarbij gaat het primair om de informatie-inhoud, niet om het behoud van oorspronkelijke formaten of structuren. De context van de data, die in de data-archiveringsliteratuur eveneens van groot belang wordt geacht, wordt zoveel mogelijk in de vorm van documentatie vastgelegd. Een belangrijk deel van het werk bestaat uit het inventariseren (en soms reconstrueren) van deze documentatie. Juist hierbij is medewerking en inzet vanuit de organisatie waarbij de archivering plaatsvindt onontbeerlijk.

Dit betekent ook dat de ADA-systematiek gebruik maakt van de zogenaamde conversie- en migratiestrategie, waarbij data leesbaar blijven door aanpassing van media en formaten aan veranderende technologische omgevingen. Om data zo weinig mogelijk te laten 'migreren' wordt doorgaans geconverteerd uitgevoerd naar een standaard, zo mogelijk software-onafhankelijk, dataformaat. In de ADA-aanpak wordt geen gebruik gemaakt van software-emulatoren om verouderde dataformaten in hun oorspronkelijke vorm te blijven aanspreken. Aanbevolen wordt wel om altijd een exemplaar van de data in het oorspronkelijke formaat te bewaren, omdat conversiefouten niet uit te sluiten zijn en documentatie te kort kan schieten. Tijdens de uitvoering van een ADA-project wordt overigens wel gebruik gemaakt van op emulatieprincipes gebaseerde software voor het bekijken en converteren van obsolete medium- en bestandsformaten.

Een probleem bij de archivering van oude digitale bestanden is dat de beschikbare informatie en documentatie over deze bestanden vaak beperkt of zelfs nonexistent is, waardoor het lastig of zelfs onmogelijk wordt om de waarde van het bestand vast te stellen. Bij voorbaat is dus niet altijd duidelijk of het bestand eigenlijk wel voor archivering geselecteerd dient te worden. Uit het ADA-project is echter gebleken – en dit is conform de ervaringen met papieren archivering volgens het herkomstbeginsel (*respect des fonds, Provenienzprinzip, principle of pro-*

venance) van archiefbescheiden – dat contextuele informatie indicaties oplevert voor de mogelijke waarde van het bestand. Dit is echter informatie van een hoger niveau, en het principe dat contextinformatie gebruikt wordt voor selectie op een lager niveau gaat op voor de hele digitale archiveringsketen. Hoe gedetailleerder wordt gewerkt, des te tijdrovender is de arbeid. Door op hoger gelegen niveaus te selecteren kan later veel werk worden bespaard.

De totale omvang van een ADA-project is afhankelijk van de omvang van de organisatie waarbij de digitale archivering plaatsvindt en de schaal waarop digitale archiefvorming heeft plaatsgevonden. De complexiteit van het project is afhankelijk van de ontwikkeling die de IT-infrastructuur heeft doorgemaakt. Het succes van het project wordt sterk beïnvloed door de bereidheid en mogelijkheid tot medewerking aan het project door de medewerkers van de organisatie zelf.

6.2 De zeven fasen van de ADA-aanpak

6.2.1 Omgevingsbeschrijving

De eerste fase kan worden beschouwd als een verkenning of een voorstudie, nodig om een verantwoord projectplan op te stellen voor de hele digitale archiveringscyclus. In deze fase wordt op het meest globale niveau de informatietechnologische infrastructuur en organisatorische context beschreven waarin de digitale dataproductie (archiefvorming) heeft plaatsgevonden. Deze beschrijving geeft een indicatie van de aard en orde van omvang van het te verwachten digitale archief. Ook worden de doelen van het project (en de gehele digitale archivering) nader gespecificeerd. Welke tijdsperiode en welke activiteiten of onderdelen van de organisatie zijn bij het project betrokken?

Resultaat	Interimrapport 1
Bronnen	Beleidsrapporten, automatiseringsplannen, gesprekken met staf en automatiserings- of ICT-medewerkers
Aanpak	Rapportanalyse, interviews
Benodigde tijd	Afhankelijk van omvang en complexiteit van de organisatie van de digitale archiefvormer: één week tot enkele maanden
Benodigd specialisme	Informatie-analyse, specialist digitale archivering (schaal 9-10)

Opzet van Interimrapport 1: Omgevingsbeschrijving
– Organisatorische context van de digitale archiefvormer
– Beschrijving van de IT-infrastructuur en de ontwikkeling daarvan
– Hardware
– Netwerken
– Software
– Gehanteerde datastandaarden
– Doelen van het digitale archiveringsproject
– Criteria voor prioriteiten
– Doelen met betrekking tot de opslag
– Doelen met betrekking tot de toegankelijkheid
– Randvoorwaarden voor de uitvoering van het digitale archiveringsproject
– Medewerking (inzet, rol) van betrokken partijen

– Stuurgroep, besluitvorming en prioriteitsvorming in het project
– Deliverables en mijlpalen
– Risicofactoren
– Aard en organisatie van de digitale informatieproductie
– Digitaliseringprocessen (organisatie van de digitalisering)
– Overzicht van digitaliseringprojecten en daarbij betrokken medewerkers

6.2.2 Materiaalafbakening

In *de tweede fase* wordt een materiaalafbakening opgesteld op het niveau van systemen en media. Welke systemen (hard- en software) zijn precies gebruikt bij de dramaproductie en op welke media zijn de gegevens opgeslagen (geweest)? Het is duidelijk dat zich hierbij al selectie- en prioriteitsvragen gaan voordoen. Het is dan ook expliciet de bedoeling dat aan het eind van deze stap is nauwkeurig is geformuleerd op welke systemen en media het vervolg van het project zich zal richten. Een selectie in hoofdlijnen (media, functiegroepen van bestanden) heeft dan al plaatsgevonden. Een voorbeeld: als in de eerste fase van het project is besloten dat het doel van het ADA-project zich beperkt tot wetenschappelijke data en dat administratief-organisatorische gegevens niet bij de archivering worden betrokken, wordt in stap 2 vastgesteld welke systemen (en versies van systemen) voor wetenschappelijk werk werden gebruikt. Puur administratieve systemen (bijvoorbeeld voor financiële en personele administratie) hoeven in deze stap (in dit voorbeeld) niet nader te worden geïnventariseerd.

Er moet ook een complete lijst komen van media waarop de data zijn opgeslagen, van mainframe tapes tot 5,25″ floppy disks tot zipdrives en backup-cassettes. Het is overigens niet altijd mogelijk om hierbij in het vervolg van het project categorieën van data uit te sluiten, omdat opslagsystemen doorgaans geen rekening houden met de aard van de informatie. Op basis van de in deze fase verworven informatie kan een globale inschatting worden gemaakt van de te verwachten problemen met de leesbaarheid en interpreteerbaarheid van de media en formaten.

Resultaat	Interimrapport 2; documentatie van de geselecteerde opslagmedia (per systeem)
Bronnen	Gesprekken met staf en automatiserings- of ICT-medewerkers, inventarisatie van gebruikte (en nog aanwezige, verouderde) computer- en opslagsystemen, inventarisatie van media en opslaglocaties
Aanpak	Interviews, inspecties on-site
Benodigde tijd	Afhankelijk van omvang en complexiteit van de computerinfrastructuur van de digitale archiefvormer: één week tot enkele maanden
Benodigd specialisme	Informatie-analist, specialist digitale archivering (schaal 9-10)

Opzet van Interimrapport 2: Materiaalafbakening
– Aanwezige media
– Bij het project te betrekken systemen
– Tijdsafbakening van bij het project te betrekken digitale collecties (lopend/actueel/afgesloten)
– Globale omvang naar type medium
– Wijze van opslag/beschikbaarheid van de media
– Beleidsaspecten en criteria voor prioriteiten

Te documenteren kenmerken van opslagmedia:
- Type: bijv. floppy disk, diskette, tape, CD-ROM, cassette, ZIP-drive, MO-drive
- Formaat: bijv. 8/5,25/3,5 inch, bandbreedte, spoelgrootte, aantal sporen
- Dichtheid: sporen/sectoren/blokken, SS/DS LD/SD/DD/HD, 800/1600/6250 BPI
- Fabrikant: relevant bij specifieke systemen, bijv. van tape streamers
- Besturingssysteem: CP/M, DOS, Windows, MVS, Unix, etc. (en versie)
- Hardware: DEC Vax, SUN Sparc, Apple IIe, CDC Cyber
- Datering: aanduiding van periode van gebruik
- Opmerkingen:

6.2.3 Selectie en inventarisatie van projecten

In *de derde fase* vindt een selectie op projectniveau plaats. Daartoe wordt een inventarisatie van archiefbestanddelen opgesteld, uiteraard alleen van die systemen en media die in het vervolg van het project worden betrokken, zoals in fase 2 vastgesteld. Onder archiefbestanddeel wordt verstaan: het geheel van archiefbescheiden (bestanden) binnen het digitale archief, bijeengebracht met een bepaald doel en in onderlinge samenhang te raadplegen. Deze te verzamelen informatie bestaat (bij archiveringsprojecten van wetenschappelijke data) overwegend uit projectinformatie en is nog niet afhankelijk van de vraag of de media en formaten van de data nog wel gelezen kunnen worden. Op grond van jaarverslagen, onderzoeksrapporten en gesprekken met onderzoekers wordt een lijst opgesteld van welke archiefbestanddelen zijn opgebouwd. Hierbij wordt ook zoveel mogelijk per project vastgesteld welke systemen en media (uit stap 2) zijn gebruikt bij het aanleggen en opslaan van de archiefbestanddelen. Ook wordt een indicatie verkregen van de omvang en homo- dan wel heterogeniteit van de bestanddelen. Hoeveel medewerkers waren bij het project betrokken? Werkte men met één of meer systemen? Hoe omvangrijk waren de aangelegde bestanddelen? Wat is de waarde voor toekomstig onderzoek? Dit zijn enkele van de vragen waarop aan het eind van stap drie een antwoord is verkregen, op grond waarvan bepaald kan worden welke projecten wel en welke niet voor digitale archivering in aanmerking komen.

Resultaat	Selectie en inventarisatie van projecten en bijbehorende opslagmedia
Bronnen	Beleidsrapporten, automatiseringsplannen, gesprekken met staf en automatiserings- of ICT-medewerkers
Aanpak	Selectie en documentatie d.m.v. rapportanalyse, interviews
Benodigde tijd	Afhankelijk van omvang en complexiteit van de organisatie van de digitale archiefvormer: één week tot enkele maanden
Benodigd specialisme	Informatie-analyse, specialist digitale archivering (schaal 9-10)

Te documenteren kenmerken van projecten:
- Naam, adres, woonplaats, etc.: NAW-gegevens van de hoofdonderzoeker/projectleider
- Werktitel project (Nederlands/Engels)
- Tijdsperiode: begin- en eindjaar van de periode waarop het project betrekking heeft
- Geografisch gebied: gebied waarop het project betrekking heeft
- Discipline/Onderzoeksthema: vakgebied en onderdeel daarvan
- Bronnen: globale aanduiding van gehanteerde bronnen (bijv. enquêtes, bevolkingsregistraties, boedelinventarissen, etc.)
- Onderwerp van onderzoek: beknopte omschrijving van het onderzoek
- Type observatie-eenheden: bijv. personen, huishoudens, etc.

– Aantal variabelen: globale aanduiding van het aantal velden
– Aantal bestanden: indien meer dan één
– Omvangindicatie: in bytes
– Aantal records: globale aanduiding van het aantal regels
– Opslagmedium/formaat
– Looptijd project: start- en einddatum van het project
– Mede-onderzoekers: namen van andere direct bij het project betrokkenen, samenwerkingspartners

6.2.4 Mediumconversie en documentatie op het niveau van dataclusters

Pas in *de vierde fase* wordt de inhoud van de geselecteerde media, systemen en archiefbestanddelen nader geïnventariseerd en beschreven in termen van groepen van bestanden, in de ADA-terminologie ook wel aangeduid als dataclusters. Deze clusters vertonen een logische samenhang en/of een inhoudelijk-organisatorische eenheid. De criteria voor wat als een cluster wordt beschouwd zijn enigszins arbitrair, maar aangezien het bij digitale informatie altijd om virtuele eenheden gaat is dit geen bezwaar. In een latere fase kunnen desgewenst aanpassingen op de clusterindeling worden aangebracht zonder consequenties voor de daarin opgeslagen informatie. Voorbeelden van clusters zijn: een projectmap (directory) op een harde schijf, een diskette, een groep bij elkaar behorende database-files die door dezelfde applicatie worden aangesproken, de CD-ROMs met images van hetzelfde project, etc. Als achteraf blijkt dat de diskette twee directories met gegevens van verschillende projecten bevat, die beter gesplitst kunnen worden in twee clusters, of dat een extra CD-ROM bij de collectie behoort, dan kan dat gebeuren zonder nadelige consequenties voor de verdere verwerking.

Bij deze stap dient ook voor het eerst aandacht besteed aan de leesbaarheid en interpretatie van de media en systemen waarop de dataclusters zijn opgeslagen. Er worden gereedschappen gebruikt om de mapstructuur en de inhoud van de mappen te kunnen lezen, die niet meer leesbaar zijn met standaard-tools. Verouderde media die verondersteld worden data clusters te bevatten worden zo nodig geconverteerd naar moderne media. Voor zover DANS zelf niet in staat is de conversie uit te voeren, wordt deze uitbesteed aan gespecialiseerde instellingen zoals het Computermuseum van de UvA.

Het is mogelijk en zelfs waarschijnlijk dat er een groep opslagmedia opduikt waarvan de inhoud bij gebrek aan enige documentatie totaal onbekend is. De hoeveelheid speurwerk die nodig is om deze media althans op clusterniveau te documenteren is niet van tevoren te ramen. Hier is doorgaans kennis en inzet van de eigenaar van de data vereist. Vervolgens wordt de selectie gemaakt van de in fase 5 te documenteren databestanden.

Resultaat	Geconverteerde/leesbare media; documentatie van de geselecteerde dataclusters
Bronnen	Geselecteerde media
Aanpak	Selectie databestanden. Conversie van oorspronkelijke media en formaten naar hedendaagse; lezen van media en directories/mappen; invoer van clusterinformatie in database
Benodigde tijd	Conversie: afhankelijk van de hoeveelheid, gangbaarheid, omvang en variëteit van de media en opslagformaten. Documentatie: afhankelijk van de hoeveelheid dataclusters
Benodigd specialisme	Conversiespecialist/systeembeheerder/computerkundige (evt. uitbesteed); data-archivist/documentalist (schaal 8-9)

Te documenteren kenmerken van dataclusters:
- Naam: naam van het cluster
- Type: map, database catalog, etc.
- Locatie: onderdeel van medium/map (padnaam)
- Grootte: in bytes
- Aantal onderliggende mappen
- Aantal bestanden
- Datum gemaakt
- Opmerkingen: relevante aanvullende informatie, bijv. conversiegeschiedenis, backups, compressie

NB: Bij ZIP-archives is aan de orde: het aantal gecomprimeerde bestanden, de compressiegrootte en de oorspronkelijke omvang, de compressiefactor, de SFX module size, etc.

6.2.5 Documentatie op bestandsniveau en conversie naar standaardformaat

In *de vijfde fase* worden de voor archivering geselecteerde bestanden gedocumenteerd op het niveau van het bestand of de database, in DANS-terminologie de dataset. Bij de selectie vindt tevens versie-analyse plaats, op grond waarvan ontdubbeling en opschoning plaatsvindt.

In eerste instantie dient de graad van detail waarop de bestandsdocumentatie plaatsvindt te worden vastgesteld. DANS hanteert voor de documentatie van bestanden een beschrijvingsmodel dat is gebaseerd op het standaard studiebeschrijvingsschema van de sociaal-wetenschappelijke data-archieven. Dit schema heeft zich ontwikkeld tot het DDI (Data Documentation Initiative) en is thans als XML schema gespecificeerd.[44]

Bij DANS worden datasets standaard beschreven in een op het DDI-schema gebaseerd systeem op het niveau van de Study, de Files en de Other Related Materials.[45] In de ADA-fasering behoort deze laatste categorie (die bijvoorbeeld ook publicaties op grond van de bestanden bevat) tot het zesde niveau. Het is noodzakelijk om in overleg met de organisatie waarbij de archivering plaats vindt de gehanteerde niveaus en de gewenste detaillering van de documentatie af te spreken.

Bij DANS worden bestanden geconverteerd naar een standaardformaat. Voor gestructureerde datasets in tabelvorm en voor tekstbestanden is dat het applicatieonafhankelijke ASCII- of XML-formaat, voor statistische bestanden het SPSS portable format. Het is ook mogelijk voor een ander standaardformaat te kiezen dat aansluit bij de huidige of toekomstige infrastructuur van de organisatie waarbij de

44 Zie: <http://www.icpsr.umich.edu/DDI/index.html>

45 Dit is het DDDI of Dutch DDI.

archivering plaatsvindt. In de nabije toekomst is archivering in het XML-formaat een optie.

Resultaat	Naar standaardformaat geconverteerde bestanden (ASCII, SPSS-portable of ander, in overleg vastgesteld formaat); bestandsdocumentatie op van tevoren vastgelegde aspecten (keuze van graad van detaillering tussen minimale en maximale variant, conform DDI)
Bronnen	Geselecteerde dataclusters
Aanpak	Conversie van oorspronkelijke bestandsformaten naar standaardformaat; documenteren van bestanden
Benodigde tijd	Conversie: afhankelijk van de hoeveelheid, omvang, variëteit en complexiteit van de data clusters. Documentatie: afhankelijk van de hoeveelheid geselecteerde databestanden
Benodigd specialisme	Conversiespecialist /data base specialist (schaal 9-10); data-archivist/ documentalist (schaal 8-9).

Minimaal te documenteren kenmerken van databestanden:
- Bestandsnaam: filenaam en extensie
- Bestandstype: bijv. Microsoft Excel Worksheet, WordPerfect 5.1 bestand, dBASE III database
- Toelichting bestandstype: in het geval van onbekende of niet gegeven extensies
- Openen met: software waarmee bestand kan worden geopend/bekeken/bewerkt
- Locatie: bijv. padnaam op schijf
- Grootte: in bytes
- Datum: hierbij kan in sommige gevallen onderscheid worden gemaakt naar de datum waarop een bestand is gecreëerd, voor de laatste keer is bewerkt/opgeslagen en voor de laatste keer is geopend/bekeken
- Auteur: naam van degene die het bestand heeft gecreëerd
- Bedrijfsnaam: naam van het bedrijf of de instelling
- Laatst opgeslagen door: naam van degene die het bestand het laatst heeft opgeslagen
- Titel: titel van document of bestand

NB: in sommige gevallen kunnen aanvullende gegevens beschikbaar zijn, zoals: onderwerp, categorie, trefwoorden, revisienummer, etc.

Maximale bestandsdocumentatie volgens het standaard DDI-schema (samenvatting):
- Document Description – This is essentially 'header' or citation information about the marked up DDI instance itself. You may decide to use only a few of the elements in this section.
- Study Description – This section describes the study at a broad level and includes information on geographic and temporal scope as well as methodological information.
- Files Description – This section is a description of the physical data file(s) in terms of record and variable counts, logical record length, etc.
- Data (Variables) Description – This section presents detailed information on each data item, including question text, variable label, category labels and values, etc.
- Other Related Materials – Other documents or files related to the study.
- Bron: <http://www.icpsr.umich.edu/DDI/users/intro-use.html>

6.2.6 Documentatie op gegevensniveau

In *de zesde fase* wordt de structuur en inhoud van ieder bestand beschreven en gedocumenteerd. Bij gestructureerde bestanden met een tabelstructuur wordt het codeboek op orde gebracht. In DDI-termen is dit de *Data Description*, waarin alle voorkomende variabelen of velden per file worden vermeld met daarbij gebruikte codes. In de praktijk van DANS wordt de informatie op het variabelenniveau doorgaans gedocumenteerd in het systeem waarin zij beschikbaar is (bijv. SPSS, codeboek als Word-bestand), omdat deze documentatiefase het meest arbeidsin-

tensief is.[46] Ook kan aanvullende informatie (zoals vragenlijsten, bronnenoverzichten, projectverslagen, gerelateerde publicaties) worden gedocumenteerd en opgeslagen, hetzij op papier, hetzij gescand als PDF-documenten.

Bij de standaardwerkwijze van DANS worden in deze fase ook controles uitgevoerd op de integriteit van de data (is de gegevensdocumentatie in overeenstemming met de bestandsinhoud en -structuur?). Indien afwijkingen worden geconstateerd (bijvoorbeeld niet-gedocumenteerde codes of velden) is het mogelijk data cleaning toe te passen (structuuraanpassingen en correcties op data en/of documentatie). Dit vergt doorgaans diepgaande kennis en analyse van de data. Het opschonen van gegevens is bijzonder arbeidsintensief en daarom kostbaar. *Data cleaning* of 'digitale restauratie' is ook niet de verantwoordelijkheid van een archief, maar die van de archiefvormer, in casu de onderzoeker.

Resultaat	Documentatie op gegevensniveau; Bestandsstructuur; Variabelenlijsten/codeboeken/DTD's
Bronnen	Geselecteerde databestanden
Aanpak	Documenteren bestanden op variabelenniveau; codeboeken op orde brengen; controle op integriteit gegevens uitvoeren
Benodigde tijd	Documentatie: afhankelijk van de aantallen databestanden en variabelen, de staat van de aangeleverde documentatie en de uitkomsten van de integriteitscontrole
Benodigd specialisme	Specialist data-archieven (schaal 9-10); data-archivist/documentalist (schaal 8-9)

Te documenteren: variabelenlijst/codeboek:
- Variabele
- Variabele-label
- Type
- Posities
- Code
- Code-label

6.2.7 Bewaring en toegankelijkheid

De toegankelijkheid van de gecreëerde metadata (of documentatie) en van de gegevensbestanden zelf wordt geregeld in *de zevende fase*. Er zijn hier verschillende opties: een instelling kan ervoor kiezen om (kopieën van) het digitale materiaal over te dragen aan DANS en een overeenkomst af te sluiten over de beschikbaarheid voor hergebruik. DANS hanteert standaard een overdrachtsovereenkomst waarin de voorwaarden voor toegang, auteursrechtelijke aspecten en de beveiliging van privacygevoelige informatie worden geregeld.

DANS hanteert verschillende toegangsniveaus, variërend van volledige publieke toegankelijkheid, via toegankelijk voor (bepaalde categorieën) onderzoekers tot ontoegankelijk voor een bepaalde duur. Aanbevolen wordt om in ieder geval een

46 Ook reconstructie en restauratie van beschadigde data en onvolledige documentatie is zeer arbeidsintensief.

kopie van de databestanden bij DANS te deponeren voor bewaring. Dit is in principe kosteloos voor de betrokken instelling. Een andere mogelijkheid is dat een instelling zowel de opslag als de toegang tot de gearchiveerde informatie geheel in eigen beheer neemt.

Resultaat	Afspraken tussen NHDA en instelling m.b.t. opslag, beheer van en toegang tot de gearchiveerde bestanden. Opslag, beheer en verstrekking van toegang tot de gearchiveerde bestanden
Bronnen	Geselecteerde en volledig gedocumenteerde databestanden
Aanpak	Overleg tussen NHDA en instelling; permanent beheer en toegang databestanden bij het NHDA of de instelling (via een website) creëren en in stand houden
Benodigde tijd	Enige tijd voor overleg nodig; afhankelijk van de grootte, complexiteit en aard der databestanden enige weken tot maanden om toegankelijkheid te verzorgen
Benodigd specialisme	Specialist data-archieven (schaal 9-10); data-archivist/documentalist (schaal 8-9)

Bijlage A

De financiële haalbaarheid van digitale archiveringsdiensten

1. Inleiding

Dit hoofdstuk gaat nader in op de financiële haalbaarheid van digitale archiveringsdiensten. In vrijwel alle publicaties van de laatste paar jaar, die aandacht besteden aan het kostenaspect van digitaal archiveren, wordt geconstateerd dat het (nog) niet mogelijk is een volledig of betrouwbaar overzicht te geven van alle mogelijke kosten, die bij langetermijnbewaring kunnen optreden. Zoals Maggie Jones en Neil Beagrie in het door hen geschreven verschenen handboek over het managen van het digitaal archiveren formuleren: '.....costs for both technical and organisational infrastructure are still not well defined'.[47]

Een van de door hen geconstateerde problemen is dat het praktisch onmogelijk is om de kosten, die nodig zijn voor het bewaren op zichzelf te scheiden van de kosten, die voor het toegankelijk maken van de data nodig zijn.[48] Meer in het algemeen gesteld spelen digitale archiveringsprojecten zich vaak in heel diverse kaders af. Dat staat nog los van feit dat daarbij de in hoofdstuk 1 gememoreerde contextverschillen een rol kunnen spelen. Een veel voorkomend verschijnsel is bij digitale archiveringsactiviteiten dat deze onderdeel zijn van een groter geheel (project of infrastructurele voorziening) waardoor het niet mogelijk is de eigenlijke archiveringskosten van andere te scheiden. Daarbij moet in ieder geval aan de overheadkosten gedacht worden van het instituut waar de digitale archiveringsprojecten zich afspelen, in het bijzonder wat betreft de IT-infrastructuur, zowel in materieel als in personeel opzicht. Digitale archivering kan ook 'meegenomen' worden in digitaliseringsprojecten. Door dit alles zijn kosten van digitale archiverings-projecten of diensten onderling moeilijk vergelijkbaar.

Zelfs wanneer de kosten van digitale archivering beperkt worden tot opslagkosten, blijft het moeilijk deze vast te stellen. Stephen Chapman publiceerde in 2003 een onderzoek naar de prijsvorming van digitale archivering. Hij vergeleek de kosten, zoals berekend aan derden, van de opslag van digitale bestanden in de

47 Jones en Beagrie (2001).
48 Jones en Beagrie (2001), 21-22.

Amerikaanse OCLC (Online Computer Library Center) met die van boeken in de universiteitsbibliotheek van de Harvard University. Hij constateerde dat in beide gevallen de opslagkosten door een aantal variabelen bepaald wordt: het overeengekomen serviceniveau, het soort depot en de wensen van de eigenaar met betrekking tot het aantal collecties, het aantal bestanden/boeken, het aantal versies en de variatie in formaten. Beslissingen over aantallen formaten en versies zijn van doorslaggevend belang voor de feitelijk betaalde prijs, aangezien deze zowel door het OCLC als door de Harvard bibliotheek op basis van de grootte van het materiaal wordt vastgesteld. Van belang hier is vooral zijn constatering dat er op dit moment geen kostenmodellen voor digitale duurzaamheid, de kosten om 'eeuwige' bewaring te kunnen garanderen, zijn ontwikkeld. Het OCLC kan nu ook geen oneindige zekerheid bieden; het vormt geen onderdeel van het contract.[49]

Het ADA-project heeft bepaalde unieke eigenschappen: het is met name gericht geweest op het achteraf archiveren van data en heeft het zich in een specifieke productieomgeving afgespeeld, die van het NIWI. Het ADA-project vertoont daardoor nog de meeste gelijkenis met het in hoofdstuk 2 vermelde Britse data-archief NDAD. Kevin Ashley (verbonden aan het NDAD) heeft enige jaren geleden in een publicatie met betrekking tot het kostenaspect van het digitaal archiveren aangegeven dat in hoofdlijnen zeventig procent van de kosten uit arbeidskosten bestaan. Dat betreft alle soorten activiteiten. De meeste tijd wordt besteed aan wat hij noemt 'depositor liaison': het contact en overleg met de opdrachtgever/beheerder van de data. Met andere woorden: het boven water krijgen, van de metadata en contextinformatie. De grootste kostenpost daarna wordt gevormd door kapitaal- en onderhoudskosten voor de hard- en software ten behoeve van de ontsluiting. Zoals door meer experts is vastgesteld: de grote kosten liggen niet zozeer in de opslag op zichzelf. Het volume van de opslag is een relatief inelastische kostenpost.[50] Zijn conclusies komen in grote lijnen overeen met die van het ADA-project.

Gelet op de bovenstaande problemen van vergelijkbaarheid en specifieke productieomgeving is het niet goed mogelijk uit het ADA pilot-project een algemeen overzicht van de kosten van de langetermijnbewaring van onderzoeksdata te distilleren. Er is daarom voor gekozen de kosten van het ADA-project hier op een zo pragmatisch mogelijke wijze weer te geven. Een overzicht wordt gegeven van de werkelijk gemaakte kosten binnen het ADA-project (A.2), gevolgd door de kosten van een specifiek deel daarvan: de bouw van de documentatie-tabel ten behoeve van de inventarisatie (A.3). Tenslotte wordt in een tabel (A.4) aangegeven welke kosten bij een toekomstig ADA-project in grote lijnen kunnen gaan optreden.

49 Chapman (2003).
50 Ashley (2000).

2 Totale kosten ADA-project

Ter toelichting moet gezegd worden dat in de tabel A.1 een onderscheid is gemaakt tussen de uitvoering van het pilot-project en de overige werkzaamheden (werkpakketten volgens het oorspronkelijke projectplan). De laatste hebben betrekking op het onderzoek gehad, evenals de overige kosten (reiskosten en aanschaf literatuur). Het pilot project heeft dientengevolge 88290,88 euro gekost (de personele kosten + de materiële kosten). Daarnaast moet benadrukt worden dat de door het Meertens Instituut gemaakte kosten hierin niet zijn opgenomen.

Tabel A.1 Kosten totale ADA-project

Kosten ADA-project, in euro's		
	Personele kosten per onderdeel	
	Opzet pilotproject	4696,63
	Verkenning (inter)nationale ontwikkelingen	19078,51
	Marktonderzoek	30840,74
	Uitvoering pilotproject	85989,31
	Opstellen eindrapport	16743,83
Totaal personele kosten		157322,00
Totaal materiële kosten		2301,57
Totaal overige kosten		1068,93
Totale projectkosten		160692,50

3 Kosten bouwen tabel voor de inventarisatie

In het volgende (tabel A.2) worden de kosten, weergegeven als uren, weergegeven van de productieve werkzaamheden van de eerste fase, dat wil zeggen de inventarisatie, de selectie en de classificering op data-clusterniveau (zie paragraaf A.3.3). Daarbij is de volgende inzet niet meegerekend:
Extern uitgevoerde dienstverlening
Ontwikkeling van procedures
Doorlopende tijdskosten: communicatie en Bitfaciliteiten
Doorlopende beheerskosten

Het materiaal is in twee subsets bewerkt. Als gevolg daarvan zijn twee tabellen ontstaan, de 'moedertabel' (a) met respectievelijk 1900 records en de complementaire clustertabel (b), op basis van 12 harde schijven, met 200 records (b).

Vervolgens is een reconstructie gemaakt. Dat wil zeggen dat een schatting is gemaakt van het aantal uren dat voor deze activiteiten benodigd is, met de kennis die wij achteraf hebben, waardoor de tabel direct kan ontstaan, zonder het vallen en opstaan van de eerste keer tijdens het ADA-project. Ontwikkelkosten, zoals het normaliseren van de tabel, zijn daarbij uiteraard niet meegerekend. Daardoor komt het aantal uren aanmerkelijk lager uit. Bij de reconstructie is rekening gehouden met de reële omvang van de dataset (bijna 1500 dataclusters), de huidige eindsituatie. Voor de hiermee corresponderende hoeveelheden zie de tabellen 2.1

Tabel A.2 Reële situatie proefproject – inzet in uren

	Activiteit	MI inzet	NHDA inzet	Totaal uren
1	opzet & bouw structuur FM-moedertabel (a)*	8	4	12
2	inventarisatie & documentatie (a)	210	21	231
3	retro aanvullende invoer & verrijking (a)	100	42	142
4	clustertabel (b) : structuur.	0	8	8
5	classificatie (a), (b)	ad 3	0	0
6	clusteren (vooral (b))	ad 3	0	0
7	inventarisatie + basisdocum. complem.tabel (b)	0	66	66
8	selectie (a), (b)	20	0	20
9a	aanvullende documentatie (2 velden; subset Brieven aan de Toekomst)*	0	3	3
9b	verrijking met 4 inhoudelijke velden (b)	ad 3	0	0
10	normalisatie moedertabel	pm	0	
			144	482

* FM = Filemaker MI = Meertens Instituut

en 3.1. Ook de mate van dienstverlening, het 'basis+'-niveau met aanvullende verrijking van 4 inhoudelijke velden, is aan de Meertens-casus ontleend

Deze berekening komt uit op een inzet van 253,8 uur, ofwel van bijna 32 werkdagen. Aan NHDA-kant liggen de werkzaamheden voornamelijk op het niveau van data-archivist (schaal 8) en IT-medewerker (schaal 6).

Tabel A.3 Reconstructie proefproject – inzet in uren

	Activiteit	(MI) inzet	NHDA inzet 8	NHDA inzet 6	Totaal uren
1	structuur clustertabel (a): 1600 losse media	0	8	0	8
2a	inventarisatie + basisdocumentatie (1550 x 3 ½)	0	0	76	76
2b	inventarisatie t/m documentatie basis+ (50 x 5 ¼)	0	0	16	16
4	structuur complementaire clustertabel	0	0	0	0
4a	facilitering inventarisatie (2; 7b)	0	8	0	8
5	classificatie	40	13,2	0	53,2
6	clusteren (ad 2 inventarisatie)	-	-	-	-
7a	inventarisatie + basisdocum. complem.tabel (b)	-	-	-	-
7b	verrijking 'basis+'niveau (2a) en (b)	0	0	66	66
8	selecteren I (clusters)	20	6,6	0	26,6
9	inhoudelijke verrijking van selectie (met 4 inhoudelijke velden)	-	-	-	-
	Totaal reconstructie 1e fase	60	35,8	158	253,8 uur

4 Kosten ADA-aanpak

Tabel A.4 geeft de kosten aan, die bij een toekomstig ADA-project in grote lijnen kunnen gaan optreden. Deze tabel volgt de indeling van de in deel II beschreven 'ADA'-aanpak, die, gebaseerd op de ervaringen van het ADA pilot-project, een

verder uitgewerkte toekomstige aanpak van digitale archiveringsprojecten biedt. In de tabel is rekening gehouden met de verschillende modules waaruit een ADA-project kan bestaan. Ook is verdisconteerd dat de bijdrage van de opdrachtgever meer of minder intensief kan zijn.

Tabel A.4 Kostenoverzicht volgens de ADA-aanpak

Activiteit	Kosten NHDA	Kosten opdrachtgever
1. Omgevingsbeschrijving	Bijna volledig arbeidskosten specialist digitale archivering)	Voornamelijk arbeidskosten: aanleveren informatie, contact en overleg
2. Materiaalafbakening	Voornamelijk arbeidskosten (specialist digitale archivering), mogelijk enige additionele onderzoekskosten (uit te besteden voor verouderde apparatuur/platforms).	Afhankelijk van de mate van cooperatie van de opdrachtgever: minimaal enige arbeidskosten voor aanleveren informatie, contact en overleg
3. Selectie en inventarisatie van projecten	Voornamelijk arbeidskosten (specialist digitale archivering)	Afhankelijk van de mate van cooperatie van de opdrachtgever: minimaal enige arbeidskosten voor aanleveren informatie, contact en overleg, speciaal i.v.m. selectie
4. Mediumconversie en documentatie data-cluster niveau	Voornamelijk arbeidskosten (IT-personeel en data-archivist en specialist digitale archivering). Server kosten	Minimaal enige arbeidskosten voor aanleveren informatie, contact en overleg, speciaal i.v.m. selectie. Documentatie afhankelijk van de mate van coöperatie van de opdrachtgever
5. Documentatie bestandsniveau en conversie	Arbeidskosten (IT-personeel, conversiespecialist, data-archivist). Server kosten	Minimaal enige arbeidskosten voor contact en overleg
6. Documentatie gegevensniveau	Arbeidskosten (data-archivist, specialist digitale archivering). Server kosten	Minimaal enige arbeidskosten voor contact en overleg
7. Bewaring en ontsluiting van de data	Arbeidskosten (data-archivist, specialist digitale archivering, website- en IT-personeel). Server kosten	Website personeel and IT-staf arbeidskosten. Server kosten
Consultancy (in alle bovengenoemde fasen)	Arbeidskosten (specialist digitale archivering)	(Arbeids)kosten voor overleg en contact

5 Conclusies financiële haalbaarheid

In dit hoofdstuk hebben wij, op grond van de ervaringen opgedaan in het ADA-proefproject (zie hoofdstuk 3) en het meer uitgebreide model voor een toekomstige werkwijze (zie hoofdstuk 6), bouwstenen aangeleverd waardoor een idee verkregen kan worden over de financiële dimensies van de activiteiten.

Een belangrijke overweging daarbij is dat de ADA-dienstverlening altijd als maatwerk zal worden aangeboden. Er is keuze uit een aantal modules mogelijk, maar vooral ook de mate waarin een opdrachtgevende instelling meer of minder zelf wil doen tijdens het project kan variëren. In het verslag van het pilot-project in het derde hoofdstuk komt dat tot uiting en ook in de bovenstaande cijfers. De toe-

komstige ADA-aanpak is er dan ook op gericht dat op zijn laatst na de derde fase daarvan (de fase 'documentatie van projecten', zie deel II) een volledig overzicht is verkregen van de hoeveelheid te verrichten werk. Vanaf dat moment is het mogelijk een verantwoorde schatting te doen van de kosten, die nog in het loop van de project kunnen optreden. Zoals in paragraaf A.3 is weergegeven komt de uitvoering van de productieve werkzaamheden van de eerste fase, dat wil zeggen de inventarisatie, de selectie en de classificering op data-clusterniveau (zie hoofdstuk 3), gebaseerd op 1500 dataclusters, neer op ruim 15.000 euro. Benadrukt moet worden dat dit niet alle werkzaamheden betreft, maar het bedrag wordt hier vermeld om van een belangrijk deel van het werkproces een indruk te geven van het financiële prijskaartje. De kosten van het opdrachtgevende instituut zelf zijn hierin niet verwerkt.

De vraag naar de financiële haalbaarheid van digitale archiveringsdiensten is niet met een simpel ja of nee te beantwoorden. Uiteindelijk zal een opdrachtgevende instelling zelf moeten bepalen of deze het bewaren van databestanden voor de lange termijn als een dermate belangrijke activiteit beschouwt dat het hierin tijd, geld en/of moeite in wil investeren. Door de gekozen systematiek (zie hoofdstuk 6) is het, samenvattend, mogelijk een toekomstig ADA-project op een dusdanige wijze uit te voeren dat de kostenrisico's beheersbaar blijven. De opdrachtgever kan echter door de modulaire aanpak en het feit dat op allerlei onderdelen ook veel eigen menskracht in een archiveringsproject gestoken kan worden, ook in dit opzicht een aanzienlijke eigen inbreng hebben.

Bijlage B

Kencijfers naar soort data

Aangeleverd				Selectie fase 1	
Classificatie-code soort data[1]	Aantal records data clusters	Getelde bestanden[2]	Kb	Bestanden	Kb
Niet leesbaar	20		1925		
DA	39	439	9976	297	10402
DM	126	1303	81168	16	827
DO	675	2216	658658	1822	248383
DP	10	52	3647	5	528
DT	107	1730	156726	785	65770
DX	20	130	25909	51	1450
PM	27	666	13968	3	1039
PS	97	2395	2283859		
PU	157	938	4570990		
PX	121	6347	297840		
SB	53	2250	219879		
overig[3]	7	14	7479		
Totaal	1459	18480[2]	8.332.031	2979	329.044

1 De classificatiecodes worden verklaard in Tabel 3.3 op pagina 24.
2 De aantallen onder 'Getelde bestanden' geven de onvolledige gegevens uit de databank weer. De onvolledigheid is het gevolg van het feit dat de inventarisatie is gebaseerd op slechts 524 van de 1460 clusters.
3 Samenvoeging van een aantal dubbelcodes.

Bijlage C

BIOM-catalogus

Datastructuur catalogus

Hieronder wordt in de tabellen C.1 en C.2 een overzicht gegeven van de veldnamen van de beide gerelateerde databanken: de cumulatieve clustertabel CCmd1 en de cumulatieve bestandentabel CBmd2.

Naar hun aard kunnen de velden in vier categorieën van beschrijvingselementen of metadata worden onderscheiden (kolom 'Categorie'):

- formeel: identificatie van cluster en bestand; ook documentaire informatie m.b.t. de elektronische locatie daarvan;
- inhoud: inhoudelijk verrijkende informatie met betrekking tot een cluster en bestand. De informatie in deze velden wordt altijd handmatig toegekend. In verband met de vereiste domeinkennis gebeurt dit door of bij de opdrachtgevende partij;
- beheer: de informatie die van belang is voor het projectbeheer, bijvoorbeeld in verband met kwantificering of de retrieval van subsets;
- technisch: een groep formele gegevens van technische aard. Heeft betrekking op zowel clusters als, vooral, op bestanden.

Met het 'ADA-niveau', in de vierde kolom, wordt bedoeld het niveau van de benodigde inzet, of anders gezegd, van de gewenste dienstverlening. De indeling speelde in het pilot-project overigens geen enkele rol, maar kan als richtlijn gebruikt worden bij toekomstige projecten. Bij de vaststelling van de mate van dienstverlening worden direct de effecten van een keus, in de vorm van de corresponderende veldenlijst, inzichtelijk. Met betrekking tot de Meertens data zijn drie niveaus onderscheiden:

1. Basis: representeert het basisniveau van de inventarisatie. De informatie van de hiertoe gerekende velden is vrijwel altijd automatisch gegenereerd.
2. Basis+ : basisinventarisatie aangevuld met enige mate van handmatige verrijking.
3. Plus: variant waarbij supplementair aan basis+ meerdere velden worden verrijkt. Inzet door of in nauwe samenwerking met de opdrachtgever.

Tabel C.1 Structuur cumulatieve clustertabel CCmd1

VeldNaam	Categorie (metadata)	Omschrijving en doel van veld	ADA - niveau	Veld type db
ada_nr	Formeel	Unieke identificatie (naam) van cluster; sleutelveld in clustertabel. Invoer: afh. van werkproces en brondata. Hier met de hand en half-automatisch gevormd.	basis	Tekst
Platform	Technisch	Besturingssysteem van bronaplicatie. Invoer: ca automatisch.	basis	Tekst
Medium	Technisch	Vorm van de gegevensdrager waarop data aangeleverd (1) Invoer: ca automatisch	basis	Tekst
TitelHmap	Formeel	Elektronische naam voor cluster. Vorming afhankelijk van gegevensdrager: – hoogste mapniveau in boomstructuur. – 'Titel'; volumelabel (en mapnaam). Zie opmerkingen hieronder.	basis	Tekst
Omschr	Inhoud	Informatie waarmee het cluster inhoudelijk wordt geduid. Losse media: ook de etiket-gegevens. Handmatige invoer van relevante informatie.	basis+	Tekst
Bestanden	Formeel/beheer	Aantal files per cluster; kan als controlewaarde worden gebruikt bij de conversie. automatisch gegenereerd.	basis	Num. (Lang)
Bytes	Formeel/beheer	Ruimtebeslag van datacluster. Aantal automatisch gegenereerd.	basis	Num. (Lang)
Kb	Formeel/beheer	Idem alternatieve Kb-equivalent bytes/1024	basis	Num. (Enkele precisie)
Srtdata	Inhoud/beheer	Door opdrachtgever toe te kennen codering cf. de classificatie. Een deel van (commerciële) software (P*) kan geautomatiseerd worden toegekend.	plus	Tekst
OpmSel	Beheer	Door opdrachtgever toe te kennen selectiecode. De waarde bepaalt of cluster doorgaat in het proces.	plus	Tekst
CatMI	Inhoud	Door opdrachtgever toe te kennen data betreffende de organisatorische indeling van het instituut / organisatie (keuzelijst). [MI = hier Meertens Instituut]	plus	Tekst
ProjMI	Inhoud	Door opdrachtgever toe te kennen data. Is een onderverdeling van CatMI, betreffende (deel)projecten (keuzelijst)	plus	Tekst
Eigenaar	Inhoud	Door opdrachtgever toe te kennen waarde. Gaat hier om de wetenschappelijk verantwoordelijke voor de data in dit cluster.	plus	Tekst
Opmerking	Inhoud	Gegevens van secundair belang (of specifiek van aard).	basis+	Tekst

Aanvullende opmerkingen bij enige veldnamen uit Tabel C.1:

Medium
Met het medium wordt hier bedoeld de gegevensdrager zoals de uitvoerder ze heeft ontvangen en bewerkt. Als het materiaal niet in de originele vorm is aangeleverd maar als kopie op een ander medium (CD), lijkt het zinvol alleen de originele gegevensdrager te vermelden. In eerste instantie is het veld bedoeld voor een eerste visuele schifting op het formaat: bijvoorbeeld diskettes 5¼″ of 3 ½″, et cetera. Indien binnen dezelfde groep media formatteringsverschillen worden geconstateerd, verdient het aanbeveling, met het oog op de benodigde apparatuur, om de tabelstructuur uit te breiden en deze technische specificaties (capaciteit, dichtheid) in een apart veld toe te voegen.

TitelHmap
Het nieuwe datamodel definieert de context van het bestand op een abstract niveau, los van het medium. Dit is ook in de tabelstructuur tot uiting gekomen, want het betrokken veld 'TitelHmap' bevat de contextinformatie van zowel losse media als die gerelateerd aan de harde schijven (directories).

1. Losse media: het Volume label of de elektronische titel van een diskette of tape. Indien de informatie op het diskette was toebedeeld aan meerdere clusters, langs de weg van de submappen, dan is aan het volume label de naam van de map toegevoegd (voorbeeld: label\mapnaam1; label\mapnaam2).
2. Data van harde schijven: de hoogste map in de boomstructuur die een datacluster identificeert. De hierbij mogelijk weggevallen pad-gegevens van de hiërarchisch lagere mappen zijn bewaard en toegevoegd als informatie op bestandsniveau ('submap').

Opmerking
Dit veld is bedoeld voor inhoudelijke informatie van secundair belang, en opmerkingen van technische aard met betrekking tot de context (bijv. foutmeldingen). De afbakening van het informatiedomein ten opzichte van het veld 'Omschrijving' dient goed geregeld te zijn..

Bytes en Kb
Deze velden kunnen beide worden gebruikt. Maar dit is afhankelijk van de wijze waarop het programma, dat de catalogisering uitvoert, deze waarden sommeert en in de uitvoer weergeeft.

Tabel C.2 Structuur cumulatieve bestandentabel CBmd2

VeldNaam	Categorie (Metadata)	Omschrijving en doel van veld	ADA - niveau	Veld type db
ada_nr	Formeel	Is sleutelveld met clustertabel. Waarden in huidige tabel niet uniek. Automatisch gevormd.	basis	Tekst
submap	Formeel	Aanvulling op >TitelHmap, vooral bij bestanden in (sub)mappen. Afhankelijk van het feit of de padnaam volledig in > TitelHmap is weergegeven. Automatisch (met handbewerking).	basis	Tekst
b_naam	Formeel	Bestandsnaam. Automatisch gegenereerd.	basis	Tekst
type	Technisch	alleen Mac-data signature: bestandstype. auto	basis	Tekst
creator	Technisch	alleen Mac-data signature: code van bron-applicatie. auto	basis	Tekst
mkdat	Technisch	datum creatie bestand. auto (via ListFiles)	basis	Datum/tijd
mktyd	Technisch	tijdstip creatie bestand. auto (via ListFiles)	basis	Datum/tijd
wzdat	Technisch	datum laatste wijziging. auto	basis	Datum/tijd
wztyd	Technisch	tijdstip laatste wijziging. auto	basis	Datum/tijd
bytes	Technisch	omvang van het bestand. auto	basis	Num. (Lang)
checksum	Beheer (technisch)	alleen Mac-data. Lokaal door besturingssysteem berekende code t.b.v. controle data integriteit van bestandsversies. auto	basis	Tekst
SEL	Beheer	Bestemd voor code van 2e selectie op het bestandsniveau. In principe handmatige invoer.	plus	Tekst

Bijlage D

Technische punten conversie

Bij het bewerken van de brondata zijn enige problemen van technische aard opgetreden, die we grotendeels hebben kunnen verhelpen. Vanwege hun algemene aard zijn ze kort het vermelden waard.

De problemen hangen samen met de niet volledige transparantie van de datainfrastructuur, en doen zich voor bij platformoverstijgend datatransport. Dit vereist enige toelichting. Het NHDA zag zich geconfronteerd met data gecreëerd op het Mac en het MS-DOS/Windows platform. Als oplossing is besloten zowel het bron- als het doelmateriaal op een derde platform op te slaan, een Novell-netwerkschijf, vanwaar de data zowel vanaf de PC als de Mac konden worden benaderd.

In de loop van het proces hebben zich daarmee problemen voorgedaan; deze waren van tweeërlei aard:

Niet volledige toegang*:* Onder Mac gemaakte bestanden bleken in een aantal gevallen niet goed benaderbaar vanaf de PC. Dit was afhankelijk van de naamgeving van de context (map) of van het bestand zelf. Concreet manifesteerde zich dit door een 'weigering' van het besturingssysteem (Windows 98) het bestand op de Novell-schijf te openen (dubbel aanklikken van file in Verkenner-venster). Hieronder (1) een voorbeeld van een directory-naam (rechts) die een probleem vormde:

(1)	cl. 0302	FM. dinn.

Naamgevingskwesties*:* Opgetreden 'spontane mutaties' in de naamgeving In BIOM komt een aantal pad- en bestandsnamen voor met gewijzigde tekens in het high-ASCII spectrum. Een voorbeeld uit BIOM met de naamweergave in de clustertabel (links) en de bestandentabel:

(2)	Dédé	DŽdŽ

Het volgende voorbeeld geeft een bestandsnaam weer via de Verkenner (links) en in BIOM:

(3)	YT204-1.23p.Ý	YT204-1.23p.¿

Ergens in het proces van inventarisatie of (geautomatiseerde) catalogisering, de daaropvolgende bewerking van de uitvoer tot proto-tabellen en het uiteindelijke inlezen hiervan in een Microsoft Access 97-databank, zijn de wijzigingen opgetreden. Ingeval van voorbeeld (2) bestaat het vermoeden dat het combineren van een Mac- en een Windows datacatalogus tot deze verschillen aanleiding heeft gegeven.

Samenvattend

- Alleen bij platformoverschrijdende databewerkingen was er een probleem
- Hoofdoorzaak daarvan waren de verschillen in naamgevingsconventie tussen de verschillende platforms
- De high-ASCII waarden in de bestandsnamen zijn niet ongevoelig voor dataverkeer over verschillende platforms (en programmatuur en de gebruikte tekensets)

Met het oog op mogelijke conflicten is het misschien wenselijk om, voorafgaande aan de conversie, de gehele bronstructuur na te lopen op dergelijke voor de PC gevoelige pad- en file-namen. De hier gebezigde vervanging met de hand zou desgewenst batch-gewijs uitgevoerd kunnen worden.

De geconstateerde technische hindernissen, die overigens al decennia bestaan, verdienen nader onderzoek, hetgeen in het kader van dit project echter te ver voerde.

Tekst encoding

De tekstbestanden zijn grotendeels op de Mac ontstaan. Na conversie bleken deze op de MAC goed leesbaar te blijven; wanneer de tekstbestanden onder Windows geopend werden was dat echter niet het geval. De teksten bleken te zijn opgeslagen onder 'Western European (Mac)' encoding, die onder Windows niet goed leesbaar is in programma's als WordPad en NotePad. Na opening in MS-Word bleek de tekst, na aanwijzing van de juiste encoding wel goed leesbaar. Indien de geopende tekst vervolgens als 'plain text' met de encoding 'Western European (Windows)' werd opgeslagen, bleek deze vervolgens ook onder Windows (inclusief Notepad etc.) goed gecodeerd te zijn.

Voorbeeld tekst encoding

Western European (Mac) encoding, zoals dit er uitziet onder Windows:

I 30-2 Soms beto™verde zÕok dÕr zeuntje. Die was nie goed bie zÕn o™d en a zÕn wee Õs wat mie zÕn uutehaele ao, lag Õn in de dune as Õn wilde te ke r te gaen.

Van oorsprong Western European (Mac), na opgeslagen te zijn als Western European (Windows):

I 30-2 Soms betoôverde z'ok d'r zeuntje. Die was nie goed bie z'n oôd `en a z'n wee 's wat mie z'n uutehaele ao, lag 'n in de dune as 'n wilde te keêr te gaen.

Literatuurlijst

Ashley K., 'Digital archive costs: Facts and fallacies', in: *Proceedings of the DLM-Forum on electronic records. European citizens and electronic information: the memory of the Information Society. Brussels, 18-19 October 1999* (Luxemburg 2000) 121-126. Op WWW: <http://europa.eu.int/ISPO/dlm/dlm99/dlm_proceed99_03.pdf>

Ashley K., 'Producing practical preservation procedures', in: *Proceedings of the DLM-Forum 2002. @ccess and preservation of electronical information: best practices and solutions* (Luxemburg 2002) 104-113

Beagrie N. en D. Greenstein, *A strategic policy framework for creating and preserving digital collections*. British Library Research and Innovation Report 107 (Londen 1998)

Bearman D., 'Reality and chimeras in the preservation of electronic records', *D-Lib Magazine* (Volume 5 Number 4, April 1999). Op WWW: <http://www.dlib.org/dlib/april99/bearman/04bearman.html>

Blauwdruk voor onderhoud, beheer en distributie van door de overheid gefinancierde digitale materialen. Publicatie Instituut voor Nederlandse Lexicologie (Leiden 2002)

Chapman S., 'Counting the costs of digital preservation: is repository storage affordable?'. *Journal of Digital Information* 4 (2003), issue 2, article 178. Op WWW: <http://jodi.tamu.edu/Articles/v04/i02/Chapman/chapman-final.pdf>

Cox R.J., *Managing records as evidence and information* (Westport, Connecticut en Londen 2001)

Dollar Ch. M., *Archival theory and information technologies. The impact of information technologies on archival practices and methods* (Universiteit van Macerata, Italië 1992)

Dollar Ch. M., *Authentic electronic records: strategies for long-term access* (Chicago 1999)

Doorn P.K. en H.D. Tjalsma: `Historical data archives: preserving and documenting historical data', in: *INSAR supplement II. The proceedings of the DLM-Forum on electronic records. Brussels 18-20 December 1996* (Luxemburg 1997) 155-160

Giesbers S., *Records Management Terminologie* (RMC Bureau; Voorburg aangevulde versie 2004). Op WWW: <http://www.rmconventie.nl/uploads/RecordsManagementTerminologiev4juli2004_1.pdf>

Gouden eieren. Notitie gebruikersgroepen en dienstverlening en inventarisatie gegevensbestanden P.J. Meertens Instituut. Interne nota van het Meertens Instituut (Amsterdam oktober 1997)

Hedstrom M., 'Electronic archives: Integrity and access in the network environment', in: S. Kenna en S. Ross, *Networking in the Humanities* (Londen, etc. 1995) 77-95

Jones M. en N. Beagrie, *Preservation management of digital materials. A handbook* (Londen 2001)

Lazinger S.S., *Digital preservation and metadata: history, theory, practice* (Englewood, Colorado 2001)

Mostert P. e.a., *Digitaal academisch erfgoed. Beleidsaspecten in verband met het behoud van wetenschappelijke digitale informatie* uitgave Surf/iWI (Utrecht 1998). Op WWW: <http://www.surf.nl/download/erfgoed.pdf>

Oltmans E. en H. van Wijngaarden, 'Digital preservation in practice : the e-depot at the Koninklijke Bibliotheek', *VINE*, 34 (2004) 21-26.

Het oog op de toekomst, Onderzoeksplan 2000-2005. Nota Meertens Instituut (Amsterdam september 1999)

Rothenberg J., *Avoiding technological quicksand: finding a viable technical foundation for digital preservation. a report to the Council on Library and Information Resources* (Washington 1999). Op WWW: <http://www.clir.org/pubs/reports/rothenberg/contents.html>

Schürer K., *Better access to electronic information for the citizen. The relationship between public administration and archives services concerning electronic documents and records management* (Luxemburg 2001).

Shepherd E. en Ch. Smith, 'The Application of ISAD(G) to the description of archival datasets', in: *Journal of the Society of Archivists*, 21 (2000), no 1, 55- 86

Shepherd E. and G. Yeo, *Managing records: a handbook of principles and practice* (Londen 2003)

Teuling A.J.M. den, *Archiefterminologie voor Nederland en Vlaanderen* (Stichting Archiefpublicaties; 's-Gravenhage 2003)

Thibodeau K., 'Knowledge and action for digital preservation: progress in the US Government', in: *Proceedings of the DLM-Forum 2002. @ccess and preservation of electronical information: best practices and solutions* (Luxemburg 2002) 175-179

Werf T. van der -Davelaar, 'Het opzetten van een digitaal depot', *Informatie Professional* 5 (2001) 20-25