Chris Bellekom en Hilde van Wijngaarden[1]

Digitale Duurzaamheid bij de KB

Digitale archivering en ontwikkeling van 'permanent access technology'

Eind vorig jaar is bij de Koninklijke Bibliotheek een digitaal archief in productie genomen. Chris Bellekom en Hilde van Wijngaarden beschrijven de activiteiten van de KB om dit digitaal archief uit te breiden met extra functionaliteit voor het langdurige beheer en de permanente toegankelijkheid van digitale documenten.

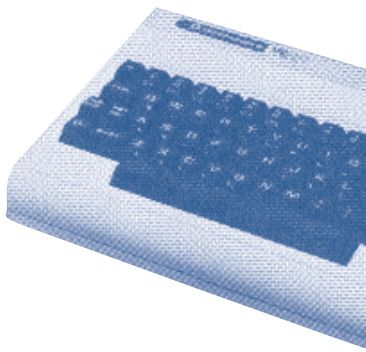
OVER DIGITALE DUURZAAMHEID wordt steeds meer geschreven. In Informatie Professional van juni dit jaar stond al een artikel van de universiteiten van Delft, Utrecht en Maastricht over een experimenteel digitaal depot.[2] Het illustreert dat het onderwerp digitale duurzaamheid steeds meer de aandacht krijgt die het wat ons betreft verdient.
Een grote hoeveelheid (wetenschappelijke) publicaties verschijnt tegenwoordig alleen nog maar in digitale vorm. De Koninklijke Bibliotheek is, als depot van Nederlandse publicaties, verantwoordelijk voor het langdurig opslaan van deze digitale publicaties en houdt zich al geruime tijd bezig met digitale duurzaamheid.
Digitale duurzaamheid bestaat uit twee hoofdonderwerpen: het zorgvuldig opslaan van digitaal materiaal (digitale archivering) en het zoeken naar methoden en technieken om digitale objecten in de toekomst nog te kunnen lezen. Deze problematiek is dermate complex dat bij de KB eind 2002 is besloten tot het oprichten van een afdeling Digitale Duurzaamheid. Daar houden twee 'digital preservation officers' zich sinds januari 2003 fulltime met dit onderwerp bezig.

Op het gebied van de digitale archivering is de afgelopen jaren veel vooruitgang geboekt. Bij de KB is in december 2002 een digitaal archief in productie genomen. Dit archief heeft de naam *e*-Depot meegekregen en is door de KB samen met IBM Nederland ontworpen en gebouwd.

Het e-Depot

Het bewustzijn groeit dat er problemen ontstaan als we niet actief werken aan het bewaren van ons digitaal materiaal. Al voor dit als probleem ervaren werd, bewaarden bibliotheken en archieven een groeiende hoeveelheid digitaal materiaal. Hierdoor werden ze al vroeg geconfronteerd met verlies van digitale informatie.
Inmiddels wordt het risico van de digitale vergankelijkheid in brede kringen erkend. Bedrijven starten projecten om de eigen digitale administratie veilig te stellen, overheden (zoals in Nederland door het Testbed Digitale Bewaring) zijn bezig met het opstellen van richtlijnen voor de bewaring van digitale stukken en universiteiten verdiepen zich in de opslag en duurzaamheid van hun wetenschappelijke onderzoeksresultaten.[3]

Het laatste decennium is hard gewerkt aan internationale afspraken en standaarden op het gebied van digitale archivering. Dat heeft onder andere geleid tot de wereldwijde acceptatie van het OAIS-model (Open Archival Information System; ISO 14721:2002) als standaard voor het opslaan van digitale objecten. Het OAIS-model is oorspronkelijk ontwikkeld door de NASA voor datamanagement. Het theoretische en generieke model is sindsdien aangepast aan de praktijk en aan de specifieke omstandigheden van bibliotheken en archieven.[4]



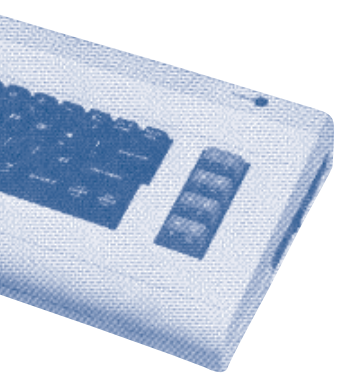
Het gebruiken van het OAIS-model als blauwdruk voor de bouw van digitale archieven is een ingewikkeld en tijdrovend proces. Instellingen die zich bezighouden met duurzaam behoud van digitaal materiaal kunnen niet à la minute hun bestaande systemen door een compleet nieuwe infrastructuur vervangen. Diverse archieven en bibliotheken zoeken naar een manier waarop de uitgangspunten van het OAIS-model ingevoerd kunnen worden in de bestaande organisatie en technische infrastructuur. Sommige instellingen, die kunnen en willen investeren in digitale opslag, werken aan een nieuw depotsysteem.
De KB heeft als eerste een compleet nieuw systeem laten ontwerpen dat gebaseerd is op het OAIS-model. Daarnaast laten ook andere grote instellingen, zoals de Duitse nationale bibliotheek en het Amerikaanse nationaal archief (NARA), een digitaal depot ontwikkelen. Ook zij nemen het OAIS-model als uitgangspunt.
Wat betreft de Nederlandse initiatieven op het gebied van digitale archivering moet het eerder aangehaalde e-Archiving project van de universiteiten van Delft, Utrecht en Maastricht genoemd worden. Het gaat hierbij om een tijdelijk project, bedoeld om voortschrijdend inzicht te verkrijgen in de problematiek rondom het duurzaam bewaren van digitale wetenschappelijke bronnen. In het digitaal archief dat gedurende dit project is ontworpen, zijn voorbeelden van verschillende soorten bronnen opgeslagen, 'ingekapseld' in XML.
Deze methode wordt ook in het buitenland onderzocht en ook al toegepast door onder andere het Nationaal Archief van Australië. De Nederlandse universiteitsbibliotheken zijn in 2003 samen met de KB en Stichting Surf gestart met het project DARE (Digital Academic Repositories).[5] Het doel van dit project is het opzetten van een infrastructuur voor de gezamenlijke opslag van de digitale academische 'output'.

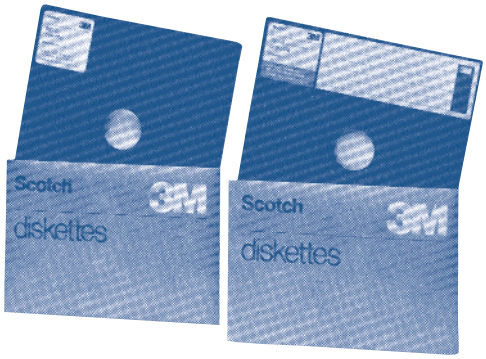
Bij de KB is in december 2002 het *e*-Depot (met als kloppend hart het IBM-systeem DIAS: Digital Information Archiving System) ingericht. Dit digitaal archief is bedoeld voor de opslag, het beheer en de beschikbaarstelling van elektronische publicaties. Het beleid van de KB is om elektronische publicaties altijd en direct na ontvangst over te brengen van het publicatiemedium (cd-rom, netwerk en dergelijke) naar de archiefomgeving.
Op 1 juli jongstleden waren er in nog geen drie maanden tijd al meer dan een miljoen elektronische artikelen geladen. Per dag kunnen ruim veertigduizend artikelen (met een gemiddelde omvang van circa 1 MB per artikel) worden toegevoegd. Na de contracten met de uitgeverijen Elsevier Science en Kluwer Academic zijn ook met andere uitgevers overeenkomsten gesloten voor permanente archivering. Binnen een jaar zal het *e*-Depot van de KB dan ook een van de grootste collecties met digitale publicaties voor de exacte wetenschappen en medicijnen in de wereld bevatten.

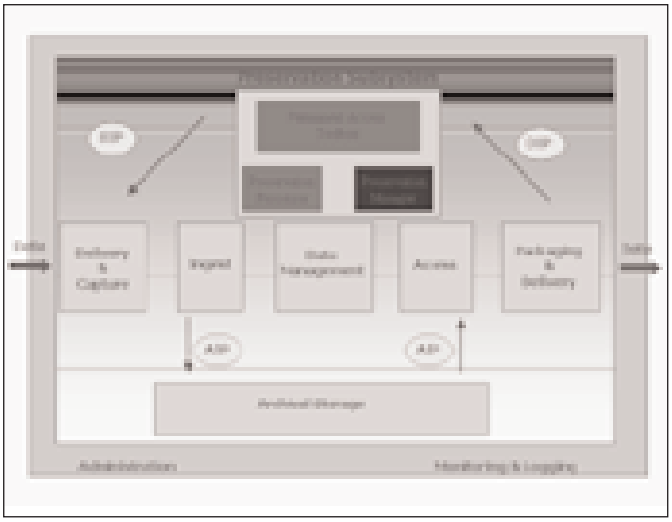
Permanente toegankelijkheid

Nu een begin is gemaakt met het zorgvuldig bewaren en beheren van het digitale materiaal, kunnen we ons gaan richten op het realiseren van permanente toegankelijkheid. Ter nadere introductie van dit onderwerp moet de terminologie even aan de orde komen.
Aangezien het hier een relatief nieuw onderzoeksgebied betreft, met een internationaal karakter, worden meestal Engelstalige termen gebruikt. Het kunnen opvragen van digitale objecten in de toekomst heet *permanent access* en wordt ook wel beschreven als *continuous rendering*. Het ontwikkelen van strategieën, procedures en tools wordt *permanent access technology* genoemd. Hoewel het inmiddels redelijk geaccepteerd is Engelstalige ICT-termen in Nederlandstalige teksten te gebruiken, zijn Nederlandse vertalingen wel mogelijk, al klinken deze nu nog wat vreemd: permanente toegankelijkheid, of gegarandeerde opvraagbaarheid.

Oplossingen voor *permanent access* komen neer op de volgende twee theoretisch meest genoemde methoden: migratie en emulatie. Bij migratie (in dit geval ook wel conversie genoemd) worden digitale bestanden geconverteerd naar een nieuw formaat. Dit kan een nieuwe versie van het originele bestandsformaat zijn, maar ook een geheel ander formaat. Conversie vraagt doorgaans relatief weinig onderzoek en ontwikkeling. Maar het risico bestaat (zeker bij een opeenvolgende reeks van conversies) dat er gegevens en functionaliteit verloren gaan.
Een van de belangrijkste alternatieven is emulatie, waarvoor veel meer onderzoek en ontwikkeling nodig is. Emulatie biedt echter wel de mogelijkheid een bestand in de toekomst op te roepen met het uiterlijk, vorm en eigenschappen zoals het bestand ooit gemaakt was; de functionaliteit en *look and feel* worden behouden. Bij emulatie wordt een configuratie van een systeem geïmiteerd op een







Figuur 1. Visualisatie van de plaatsing van het Preservation Subsystem binnen het OAIS-model

ander systeem, of wordt een los programma gesimuleerd. Een emulator kan het best gemaakt worden als het te emuleren systeem of programma nog werkt. Voor ieder nieuw systeem moet dit opnieuw gebeuren. Dit maakt emulatie een arbeidsintensieve oplossing, maar tegelijk technologisch een zeer uitdagende strategie.

Bij de keuze voor een strategie is het belangrijk te bedenken wat we in de toekomst, over vijftig tot honderd jaar, met het opgeslagen materiaal willen kunnen doen. Als we de inhoud ongeveer willen behouden, is het wellicht voldoende om het bestand telkens te migreren naar nieuwe formaten. Veranderingen in uiterlijk, functionaliteit of in details van de inhoud, zijn dan minder belangrijk.
Als we in de toekomst ook willen weten hoe een digitaal object er oorspronkelijk uitzag, kunnen we volstaan met het bewaren van een afbeelding ervan. Maar als we willen dat het document in zijn originele vorm, met de functionaliteit en de integrale inhoud van het moment van opslag over vijftig jaar op te vragen is, stelt dit andere eisen aan de *permanent access technology*. Dan lijken op emulatie gebaseerde oplossingen het beste.
Het is ook mogelijk dat we in de toekomst documenten willen bekijken en gebruiken met extra functionaliteit die op dat moment binnen bereik is gekomen: vormen van zoeken, bewerken en presenteren (op een wijze die eerder niet mogelijk waren) met een snelheid en kwaliteit die op dat moment gangbaar is. Dat vereist nog weer een andere aanpak, een waarbij data-export mogelijk is naar de dan gangbare gebruikersomgeving.

Onderzoek naar permanente toegankelijkheid van digitale objecten is tot nu toe alleen kleinschalig en door individuele organisaties uitgevoerd. Er zijn nog geen internationale afspraken gemaakt, of standaarden ontworpen, zoals voor digitale archieven. In verschillende landen zijn er enkele veelbelovende onderzoeken uitgevoerd en prototypes getest, maar de technologische ontwikkeling en internationale samenwerkingsprojecten staan nog in de kinderschoenen.[6]
In Nederland is veel onderzoek gedaan bij het Testbed Digitale Bewaring. Dit team heeft enkele interessante testen uitgevoerd met migratie en emulatie.[7] *Permanent access* is ook bij het e-Archivingproject aan de orde geweest. Zij hebben een methode ontwikkeld waarbij het opgeslagen digitale materiaal bekeken kan worden met opgeslagen viewers die draaien op een centrale computerserver. Een operationele oplossing voor toekomstige toegankelijkheid van digitale documenten is op dit moment niet beschikbaar. Wel is duidelijk dat één oplossing niet voldoende zal zijn: bij permanente toegankelijkheid hangt de oplossing zowel af van het doel waartoe iets bewaard wordt, als van het soort digitaal object.

De Preservation Manager

Om het langdurig behoud van en de permanente toegankelijkheid tot de elektronische publicaties in het *e*-Depot te ontwikkelen, is een team van de KB samen met IBM begonnen met het bouwen aan een zogenaamd 'Preservation Subsystem' (zie figuur 1). Dit subsysteem zal bestaan uit een Preservation Manager, een Preservation Processor en 'tools' voor *permanent access*.
De Preservation Manager beheert, controleert en signaleert met behulp van technische metadata de integriteitsstatus van de digitale objecten die zijn opgeslagen in het *e*-Depot. De Preservation Processor wordt gestart als de integriteit of toegankelijkheid in gevaar komt. Deze functie zal documenten aanbieden voor bewerking en aanpaste versies of kopieën opnieuw laden. KB en IBM zullen ook een operationele tool voor permanente toegankelijkheid ontwikkelen: de zogenaamde Universal Virtual Computer (UVC) (zie figuur 2) voor image-formaten.

Documenten in het *e*-Depot hebben bepaalde *file formats* (voor e-journals is dat voornamelijk PDF). De Preservation Manager registreert de *file formats* in het *e*-Depot.[8] Door de Preservation Manager worden deze technische metadata in een database in XML opgeslagen. Daarin staat beschreven welke applicatie nodig is om een document te openen, welk operating system vereist is voor de applicatie en op welke processor het operating system draait. Zo wordt voor ieder file format een pad (View Path) in het *e*-Depot vastgelegd waarmee de documenten weergegeven kunnen worden. Voor één enkel file format zijn meerdere paden mogelijk, die alle geregistreerd worden. Mocht een van de elementen van een pad niet meer beschikbaar zijn, dan is het document altijd nog via een ander pad toegankelijk. Naarmate de tijd vordert zullen steeds minder paden beschikbaar zijn. De Preservation Manager lost dat probleem weliswaar niet op, maar signaleert deze, zodat actie kan worden ondernomen. Een van de acties die ondernomen kan worden als een bepaald file formaat niet meer beschikbaar is, wordt hierna beschreven.

Een 'Universal Virtual Computer'

Zoals eerder beschreven, is er nog lang niet genoeg onderzoek gedaan naar de verschillende (technische) aspecten van *permanent access*. Ook zal één enkele oplossing niet volstaan. Het IBM/KB-team is gestart met het operationaliseren van één mogelijke *permanent access tool*: een uitwerking van het door IBM-expert Raymond Lorie ontwikkelde Universal Virtual Computer (UVC) concept.[9] Door het enkel in

theorie bestaande UVC-concept in de praktijk te ontwikkelen, kan het verdere onderzoek beter gericht worden.

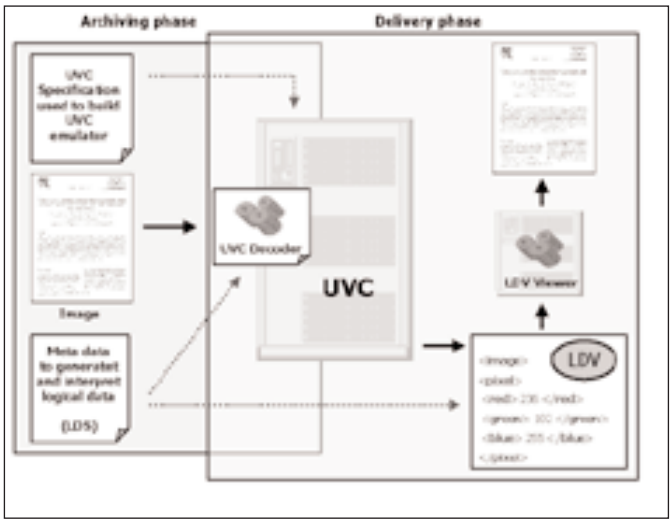
Een UVC is een zo eenvoudig mogelijk programma dat kan draaien op ieder mogelijk nieuw computerplatform. Een voorbeeld van zo'n eenvoudig programma is Ms-dos: begonnen als besturingssysteem in 1981 en nog steeds als applicatie draaiend in 2003. Voor een specifiek formaat wordt een voor de UVC geschreven applicatie gemaakt. Dit is een 'vertaler' (Decoder) waarmee het bestand wordt omgezet naar een zogenaamde logische data view (LDV). Een LDV is een gestructureerde beschrijving van het object, opgesteld volgens een schema (logical data schema of LDS). Stel dat iemand over vijftig jaar iets wil opvragen dat in 2003 is opgeslagen, dan kunnen programmeurs de UVC – immers zelf ook een programma – aanpassen voor de computer van 2053. Vervolgens kan de Decoder worden gestart en de LDV gegenereerd. Omdat behalve de Decoder destijds ook de LDS is gearchiveerd, kunnen 'de programmeurs van de toekomst' een viewer maken. Het betreffende document kan daarmee bekeken worden en heeft dan dezelfde functionaliteit en hetzelfde uiterlijk als het opgeslagen origineel.

Omdat het realiseren van een operationele UVC voor een formaat als PDF op dit moment nog zeer veel ontwikkeling vraagt, begint het KB/IBM team eerst met een UVC voor image formaten. Een image is eenvoudiger te analyseren dan een PDF-bestand, waardoor het mogelijk is er een Decoder voor te ontwikkelen. Het resultaat is dan, dat een digitaal object in het *e*-Depot bij opvraag geconverteerd kan worden naar een image-formaat (jpeg, png of tiff). Voor dit image formaat zal een logisch data schema (LDS) en een Decoder zijn ontwikkeld die ervoor zorgen dat de afbeelding altijd te bekijken zal zijn.

De UVC voor images zal begin 2004 volledig ontwikkeld zijn, inclusief de componenten die bij deze strategie pas in de toekomst gebruikt zullen worden. Het gaat hier om een toepassing die alleen aangewend zal worden als bestandsformaten in het *e*-Depot niet meer via hun eigenlijke viewers te bekijken zijn. Bij deze strategie hebben we alleen nog een afbeelding van een artikel: leesbaar en in de originele opmaak, maar niet te doorzoeken of te bewerken. Deze procedure heeft dus beperkingen, maar zal wel de eerste werkende oplossing voor permanente toegankelijkheid zijn. Het biedt de garantie dat we alles wat we nu opslaan altijd zullen kunnen bekijken.

Internationaal programma

Als we digitale objecten zorgvuldig hebben opgeslagen, kunnen we niet passief afwachten of we deze in de toekomst nog kunnen bekijken. Er simpelweg op gokken dat iemand in de toekomst oude bestandsformaten zal kunnen emuleren zonder dat we daar nu de voorwaarden voor creëren, is niet genoeg. Zo is voor het schrijven van een emulator informatie over eigenschappen van computersystemen nodig, die wellicht in de toekomst niet meer (of slechts met de grootst mogelijke moeite en dus kosten) te achterhalen zijn.
Voor de KB, als depotbibliotheek met de verantwoordelijkheid publicaties voor de lange termijn te bewaren, is het



Figuur 2. Overzicht van het process om images in de toekomst te kunnen opvragenvia een Universal Virtual Computer

belangrijk dat we de ontwikkeling van *permanent access* technologie stimuleren. In de ICT-industrie krijgt het probleem van langdurige houdbaarheid immers nog onvoldoende prioriteit. Daarom werken wij, met een groep andere instellingen, aan het ontwikkelen van een internationaal programma voor verder onderzoek en standaardisering van permanente toegankelijkheid van digitale informatie.

Noten

1. Met dank aan Johan Steenbakkers en Ingeborg Verheul.
2. Ronald Dekker en Martin Slabbertje, 'E-Archiving. Het duurzaam bewaren van wetenschappelijke digitale bronnen', *Informatie Professional*, juni 2003, jrg. 7, no. 6, p. 32-34.
3. Zie voor enkele Nederlandse initiatieven op het gebied van digitale duurzaamheid: www.digitaleduurzaamheid.nl en www.library.tudelft.nl/e-archive/, www.digidiv.nl, www.surf.nl/themas/index2.php?oid=18.
4. De partners in het Europese project NEDLIB (Networked European Deposit LIBrary, 1998-2000) hebben het OAIS-model gebruikt voor het opstellen van een model voor digitale archivering door bibliotheken. Zie o.a. Titia van de Werf, *A Process Model: The Deposit System for Electronic Publications* (NEDLIB Report series 6), Den Haag 2000.
5. Zie www.surf.nl/themas/index2.php?oid=18.
6. Een beschrijving van de internationale ontwikkelingen op het gebied van *permanent access technology* zou in dit artikel te ver voeren.Voor wie hier meer over wil lezen, biedt de internationale kennissite PADI: www.nla.gov.au/padi toegang tot diverse publicaties en websites.
7. Zie voor de diverse projecten van het 'Testbed': www.digitaleduurzaamheid.nl.
8. Het ontwerp is gebaseerd op: R.J. van Diessen, *Preservation requirements in a deposit system IBM/KB long-term preservation study report series number 3* (2002).
9. Onder andere beschreven in: R. Lorie, *The UVC, a method for preserving digital documents: proof of concept IBM/KB long-term preservation study report series number 4* (2002).

Voor meer informatie over de lopende projecten bij de KB: www.kb.nl/e-depot/

Chris Bellekom en Hilde van Wijngaarden zijn beide digital preservation officer bij de afdeling Research & Development van de Koninklijke Bibliotheek.