

Digitaal archiveren en digitale duurzaamheid

Filip Boudrez
Expertisecentrum DAVID vzw
Antwerpen, 2009

0. INHOUD

1. Inleiding.....	1
2. Een risicovol reconstructieproces.....	1
3. Digitale objecten.....	3
4. Digitale documenten.....	5
5. Digitale archiefdocumenten.....	9
6. Besluit.....	11

1. INLEIDING

Het archiveren van digitale informatie verschilt in een aantal opzichten fundamenteel van het bewaren en het raadplegen van informatie in analoge vorm. Veel van die verschillen vinden hun oorsprong in WAT in de analoge en de digitale wereld wordt gepreserveerd en wordt geraadpleegd. In de analoge wereld vormen de gegevensdrager en het document een fysieke eenheid. Hierdoor geldt bijv. voor papieren documenten of voor 8/16 mm. films dat wat je bewaart en raadpleegt identiek hetzelfde is, nl. het document. De belangrijkste voorwaarde om een analogoog document in tijd over te brengen, is de preservering van zijn gegevensdrager in goede materiële staat. In de digitale wereld is dit niet zo. Van een digitaal document bewaren we de bits en bytes. Die bits en bytes vormen doorgaans geen fysieke eenheid met hun opslagmedium. Wel integendeel, computerbestanden kunnen relatief gemakkelijk van de ene gegevensdrager naar de andere worden overgezet zonder dat dit merkbaar is of met informatie- en kwaliteitsverlies hoeft samen te gaan. Belangrijker is evenwel dat de bewaring van digitale gegevensdragers in goede staat geen enkele garantie biedt dat digitale archiefdocumenten leesbaar en bruikbaar blijven.

Ook het raadplegen van analoge en digitale informatie verschilt grondig. Wat je in de digitale wereld raadpleegt, is een document dat telkens opnieuw moet worden opgebouwd op basis van die opgeslagen bits en bytes. Voor het raadplegen en het reconstrueren van digitale informatie is technologie of een combinatie van hard- en software nodig. Bits en bytes worden opgeslagen op gegevensdragers waarvoor leesapparaten, interfaces en stuurprogramma's nodig zijn. Digitale documenten zijn opgeslagen in een bepaald bestandsformaat en kunnen dikwijls enkel met specifieke software of zelfs enkel met welbepaalde versies van die software worden gelezen. Stuurprogramma's en applicatiesoftware zijn op hun beurt dan weer afhankelijk van besturingssystemen, terwijl die besturingssystemen specifieke hardwarecomponenten nodig hebben. Hard- en software evolueren echter aan een razendsnel tempo, zonder dat digitale informatie op oudere generaties gegevensdragers of in oude bestandsformaten op nieuwere computersystemen leesbaar blijft. De achterwaartse compatibiliteit is immers in de regel beperkt tot hooguit een klein aantal generaties. Er is geen enkele garantie dat digitale documenten gecreëerd met de technologie van vandaag, raadpleegbaar zijn met toekomstige technologieën.

Samen met de grote impact van fouten of anomalieën is die afhankelijkheid van hard- en software er verantwoordelijk voor dat digitale archivering heel kwetsbaar is. Digitale informatie overleeft minder gemakkelijk de tand des tijds dan analoge informatie. Digitale archivering is evenwel niet onmogelijk, maar vraagt een continue zorg en een proactief beheer waarbij de nodige maatregelen worden getroffen om digitale documenten leesbaar en authentiek te houden.

2. EEN RISICOVOL RECONSTRUCTIEPROCES

Digitale gegevensdragers bevatten digitale objecten, maar gebruikers raadplegen digitale documenten. Een digitaal archiefdocument moet bij raadpleging als het ware telkens opnieuw worden gereconstrueerd op basis van de gepreserveerde bits en bytes¹. Hoe een digitaal document er finaal op scherm uitziet, is afhankelijk van de hard- en software waarover de gebruiker beschikt en van zijn/haar voorkeursinstellingen². Digitale archivering als een reconstructieproces beschouwen, biedt een goed perspectief om zicht te krijgen op de functionaliteiten en noden van een goede digitale archiveringsoplossing. Het einddoel van dit proces is de bewaring van documenten die leesbaar, begrijpbaar en bruikbaar zijn voor mens en machine. Dit veronderstelt de duurzame bewaring van:

- digitale objecten: de opgeslagen bits en bytes zijn intact en kunnen naar het computergeheugen worden getransfereerd
- digitale documenten: de objecten zijn opgeslagen in een ondersteund bestandsformaat zodat het document op menselijk leesbare wijze kan worden gepresenteerd
- digitale archiefdocumenten: de gearchiveerde documenten zijn bruikbaar, wat impliceert dat ze vindbaar, begrijpbaar, authentiek en betrouwbaar zijn.

De benadering van digitaal archiveren als een reconstructieproces vestigt de aandacht op wat in de toekomst nodig is om digitale archieven te raadplegen. De moeilijkheid hierbij is dat de digitale archiefbeheerder niet weet hoe de digitale toekomst er zal uitzien. Digitaal archiveren kan dan ook vergeleken worden met het samenstellen van een digitale tijdscapsule. De inhoud van die tijdscapsule moet raadpleegbaar zijn op computersystemen waarvan we niet weten hoe die zullen functioneren of welke gegevensdragers en bestandsformaten ze ondersteunen. Digitaal archiveren is het archiveren van de mogelijkheid tot reconstructie. Dit kan betekenen dat niet alleen de digitale archiefdocumenten worden gearchiveerd, maar ook de digitale componenten en de documentatie die bij raadpleging van de digitale archieven nodig zijn.

Koppelt men deze benadering van digitale archivering aan risico-analyse, dan kunnen onmiddellijk al twee belangrijke conclusies worden geformuleerd. Ten eerste, hoe meer schakels het reconstructieproces bevat, des te kwetsbaarder of risicovoller het proces is. Van zodra één schakel in het reconstructieproces ontbreekt, dient het digitaal archiefdocument als verloren te worden beschouwd. Voor een veilig en bedrijfszeker digitaal archiveringsproces is het bijgevolg aangewezen om zo weinig mogelijk schakels te hebben. Vermijdbare of overtollige reconstructieschakels worden beter niet toegepast, want deze brengen extra afhankelijkheden en risico's met zich mee. Daarom is het gebruik van compressie, encryptie, paswoorden, enz. beter te vermijden. Niet alle reconstructieschakels kunnen echter worden vermeden. Finaal zal nog steeds een combinatie van hard- en software nodig zijn voor de raadpleging. De tweede conclusie is dan ook dat voor de overblijvende schakels de archiefbeheerder bij voorkeur zoveel mogelijk risicospreiding toepast, zodat afhankelijkheden ten aanzien van specifieke producenten, software(versies), hardwarecomponenten worden vermeden. Vanuit die optiek is het belangrijk om voor de overblijvende reconstructieschakels zoveel mogelijk normen toe te passen. Immers, de (technische) specificaties van normen zijn gedocumenteerd en normen worden doorgaans door verschillende technologieën of producenten ondersteund. Bij het zoeken naar oplossingen voor de drie vermelde deelaspecten van digitale archivering past de beheerder van digitale informatie beide conclusies best zo consequent mogelijk toe.

¹ Deze zienswijze is o.a. uitgewerkt door de preservation task force van InterPARES 1 (Preservation Task Force, *How to preserve authentic electronic records*, 2001).

² Het Nationaal Archief van Australië hanteert voor deze visie het *performance* model: net zoals iedere opvoering van hetzelfde toneel- of muziekstuk op een aantal punten verschilt, kan het gereconstrueerde digitaal document afwijken van het 'origineel'. NATIONAL ARCHIVES OF AUSTRALIA, *An approach to the preservation of digital records*, Canberra, 2002.

Samen met de andere uitgangspunten van digitaal archiveren leidt dit tot vijf uitgangspunten bij het bouwen en beheren van een digitaal archief:

1. het vermijden van (externe) afhankelijkheden
2. het vermijden van overbodige reconstructiestappen
3. het toepassen van risicospreiding en -beheer
4. het gebruiken van normen en standaarden
5. het documenteren van de implementatie en het beheer.

Uit bovenstaande blijkt duidelijk dat het niet volstaat om alleen de bits en bytes van digitale objecten of de gegevensdrager op een duurzame wijze te bewaren. Wil men digitale informatie archiveren, dan zijn ook oplossingen nodig om de leesbaarheid, begrijpbaarheid en betrouwbaarheid te bewaren. Het conserveringsbeleid mag zich bijgevolg niet alleen richten op het conserveren van de gegevensdragers en hun inhoud.

3. DIGITALE OBJECTEN

Elke oplossing voor digitale archivering vereist dat de digitale objecten intact en inleesbaar worden bewaard. Bits en bytes hebben echter de neiging om "omver te vallen", waardoor ze niet meer correct kunnen worden ingelezen. Dit verschijnsel wordt 'bit rot' genoemd en kan afhankelijk van het type gegevensdrager verschillende oorzaken hebben. Bij harde schijven of flash memory³ is dit doorgaans een afzwakking van de elektrische lading waardoor positieve neutronen veranderen in negatieve. Bij optische dragers zoals CD's of DVD's kan corrosie van de reflecterende metalen laag tot een daling van de straalintensiteit leiden. Tapes zijn dan weer niet enkel het voorwerp van een afname van magnetisme, ze zijn daarenboven sterk onderhevig aan gebruiksslijtage. Dit kan leiden tot corrupte of beschadigde computerbestanden. De oorzaken beperken zich overigens niet alleen tot degeneratie en/of beschadiging van de gegevensdrager, maar bugs in hard- en software kunnen evengoed aan de basis liggen.

Kortom, fouten zijn inherent aan digitale opslag. Elke digitale gegevensdrager bevat fouten, maar de eindgebruiker merkt hier doorgaans weinig of niks van. De hard- en software voor het lezen van gegevensdragers is immers standaard met automatische foutopsporing en -verbetering uitgerust. Tot een bepaald kritisch punt kunnen fouten in de gegevensopslag automatisch worden hersteld. Alleen wanneer het aantal fouten dit kritisch punt overstijgt, zijn de digitale objecten onherstelbaar beschadigd en dient bijgevolg het digitaal archiefdocument als verloren te worden beschouwd.

Om digitaal gegevensverlies te voorkomen of te herstellen is het aangewezen om voortdurend de integriteit van de gearchiveerde bits en bytes te controleren. Het mechanisme dat hiervoor algemeen wordt gebruikt is het vergelijken van CRC's⁴ of van checksums⁵. Bij archivering dienen hiervoor de correcte CRC's of checksums van de digitale objecten te worden berekend en geregistreerd. Door de CRC of checksum te herberekenen en te vergelijken met de oorspronkelijke waarde kunnen fouten in de bitopslag worden opgespoord. Idealiter is deze vorm van kwaliteitscontrole een permanent en systematisch proces.

³ Flash memory is het type gegevensdrager dat wordt gebruikt in USB-sticks en geheugenkaarten voor digitale camera's, PDA's, MP3-spelers, GSM's en smartphones.

⁴ CRC (Cyclic Redundancy Check) is specifiek ontworpen om fouten in de bitopslag op te sporen. CRC is geïmplementeerd in een aantal protocollen en een aantal bestandsformaten zoals ZIP, TAR en PNG. CRC's worden ook standaard gebruikt door besturingssystemen om computerbestanden correct en gecontroleerd te kopiëren.

⁵ Een checksum kan als het ware vergeleken worden met een digitale vingerafdruk van een digitaal object. Elk digitaal object heeft in principe een unieke checksum, al is dit ook afhankelijk van de sterkte van het checksumalgoritme. Het sterker het checksumalgoritme, des te meer rekenkracht het berekenen vraagt. Veel gebruikte checksumalgoritmes zijn MD5, SHA-1 en SHA-2.

Het opsporen van fouten in de bitopslag is echter maar de eerste stap. De fouten moeten ook kunnen worden verbeterd of hersteld. Hiervoor zijn pariteitsgegevens, reservekopieën en back-ups nodig. Samen met het opsporen van eventuele fouten in de bitopslag vormen herstel of verbetering de kern van elke beheersprocedure voor een correcte digitale gegevensopslag. Bij gegevensdragers bestemd voor gewoon consumentengebruik zoals (externe) harde schijven, CD of DVD zijn beide handelingen een vrij arbeidsintensief proces, want automatisering van deze handelingen is niet altijd mogelijk. Desondanks kunnen deze types opslagmedia voor archivering worden gebruikt, als de nodige voorzorgs- en beheersmaatregelen worden getroffen⁶. Deze soorten gegevensdragers hebben als voordeel dat ze een lage investeringskost hebben en nagenoeg geen gespecialiseerde kennis vergen.

Doordat de courante consumenten opslagmedia nog veel handmatige handelingen vragen, zijn zij niet geschikt voor de archivering van heel grote volumes aan digitale objecten. Bovendien zijn bij deze opslagmedia slechts periodieke of steekproefgewijze kwaliteitscontroles mogelijk. Meer geavanceerde opslagsystemen zoals NAS-boxen (NAS: Network Attached Storage⁷) of SAN's (Storage Area Network⁸) houden wel automatisch een bepaalde hoeveelheid pariteitsgegevens bij en voeren wel continu kwaliteitscontroles uit. Die kwaliteitscontroles richten zich op de performantie van de harde schijven en de bitintegriteit van de digitale objecten. Samen met hun grote opslagcapaciteit worden dergelijke opslagsystemen hierdoor als meer robuust en meer geschikt voor langetermijnarchivering beschouwd.

De hoeveelheid pariteitsinformatie die deze systemen bijhouden is afhankelijk van het type RAID-configuratie⁹ dat wordt gehanteerd. RAID-5, RAID-5+ en RAID-6 worden in de praktijk veel toegepast. Een NAS of een SAN vragen in tegenstelling tot de courante opslagmedia wel een aanzienlijke investering, al zijn NAS-boxen de voorbije tijd opmerkelijk goedkoper geworden zodat hun aanschaf wel financieel haalbaar is voor kleine of middelgrote organisaties, of zelfs particulieren.

NAS- of SAN-opslagsystemen dekken meer risico's af dan gewone opslagmedia, maar zijn evenmin foutenvrij. Recent onderzoek naar deze geavanceerdere opslagsystemen wees uit dat deze helemaal niet onfeilbaar zijn en dat ze ook corrupte digitale objecten kunnen bevatten¹⁰. Een NAS of een SAN kan onderhevig zijn aan 'silent data corruption' of 'silent disk corruption' zonder dat dit wordt vastgesteld door de beheerssoftware van het opslagsysteem. Om die reden is het geen overbodige luxe om toch nog expliciet CRC's of checksums te registreren en op basis daarvan nog afzonderlijke kwaliteitscontroles uit te voeren. Aangezien alle data op een NAS of SAN online toegankelijk zijn, kan deze kwaliteitscontrole wel volledig automatisch worden uitgevoerd. Fouten herstellen dient dan op basis van een back-up te gebeuren. Back-ups zijn een essentieel onderdeel van de beheersprocedure bij het gebruik van een NAS of SAN, maar gelet op de grote omvang van digitale archieven kunnen niet zomaar standaard back-upregimes worden toegepast. Back-ups dienen overigens niet alleen om

⁶ Het DAVID-project publiceerde twee praktische richtlijnen voor het gebruiken van optische en magnetische dragers voor archiveringsdoeleinden. Zie:

- *Digitaal ArchiVeren: richtlijn en advies*, nr. 2: *Duurzame CD's* (<http://www.edavid.be/davidproject/teksten/Richtlijn2.pdf>)
- *Digitaal ArchiVeren: richtlijn en advies*, nr. 6: *Duurzame magnetische dragers voor het archief* (<http://www.edavid.be/davidproject/teksten/Richtlijn6.pdf>)

⁷ Een NAS is een opslagsysteem dat aan een computer of een netwerk is gekoppeld. Een NAS is met een besturingssysteem uitgerust. Dit besturingssysteem bepaalt in welk bestandssysteem de documenten worden opgeslagen.

⁸ Een SAN is een opslagsysteem zonder eigen besturingssysteem en dat is gescheiden van de servers waarmee ze worden beheerd. De aansturing gebeurt vanuit deze servers. Het bestandssysteem dat een SAN gebruikt, is dat van de servers.

⁹ RAID (Redundant Array of Independent Disks) is een configuratie waarbij meerdere harde schijven met elkaar worden gecombineerd en waarbij de digitale objecten en hun pariteitsinformatie voortdurend over meerdere schijven wordt gespreid.

¹⁰ B. PANZER-STEINDEL, *Data integrity*, 2007.

systeemfouten te herstellen. Beheerders van digitale archieven kunnen ook wel eens de verkeerde digitale objecten wissen.

Vanwege deze laatste reden is het 'mirroren' of 'spiegelen' van de digitale objecten in een tweede opslagsysteem geen afdoende veiligheidsmaatregel. Een 'mirror' is een parallel opslagsysteem dat wordt gebruikt voor het verzekeren van de performantie (load-balancing) en als extra veiligheidsmaatregel in geval van calamiteiten zoals een systeemcrash, een brand, een overstroming, enz. De gegevens worden tussen beide opslagsystemen met een bepaald interval gesynchroniseerd zodat eventuele vernietigingen op systeem 1 ook op systeem 2 worden doorgevoerd. Met andere woorden, gewisse digitale objecten zijn niet recupereerbaar in het parallelle opslagsysteem. Hiervoor zijn back-ups nodig of men dient ervoor te zorgen dat digitale objecten wel logisch maar niet fysiek worden verwijderd.

In geval van opslag op harde schijven is een 'mirror' wel een belangrijk instrument om de risico's te spreiden. Naast de hierboven vermelde calamiteiten biedt een 'mirror' ook de mogelijkheid om de digitale objecten in verschillende bestandssystemen bij te houden.

Het intact en veilig preserveren van bits en bytes is slechts één basisvereiste om succesvol digitale objecten te archiveren. De gearchiveerde bits en bytes moeten ook door toekomstige computersystemen kunnen worden ingelezen. Dit vereist niet alleen dat de nodige leesapparatuur en -programmatuur aanwezig zijn, maar ook compatibel zijn met de computersystemen waarmee de raadpleging gebeurt. Stuurprogramma's voor de gegevensdragers en de toegepaste bestandsindeling moeten compatibel zijn met besturingssystemen en hun bestandssystemen. Van zodra op één van deze factoren de vereiste ondersteuning dreigt weg te vallen, is het aangewezen om de digitale objecten over te zetten naar een nieuw type gegevensdrager die wel nog volop wordt ondersteund.

Als algemeen uitgangspunt wordt aangenomen dat gegevensdragers mits een goede beheersprocedure wel lange tijd intacte digitale objecten kunnen bevatten, maar dat de leestechnologie veel sneller in onbruik geraakt. Alleen nog maar vanwege deze reden moet de inhoud van digitale gegevensdragers vrij frequent worden overgezet. Voor audiovisuele archieven worden bijgevolg specifieke tapeformaten best niet gebruikt. Beter is om een dataformaat te gebruiken dat autonoom is ten aanzien van zijn opslagmedium. Digitaal geluid wordt bijvoorbeeld beter gearchiveerd als WAVE-bestand dan als audioCD of als MiniDisc. Digitale video wordt beter gearchiveerd als MXF-bestand dan als digitale betacamtape.

Afhankelijkheden ten aanzien van specifieke leveranciers of technologieën worden bij de keuze van gegevensdragers best zoveel mogelijk vermeden. Absolute garanties voor langetermijnleesbaarheid zijn er niet, maar door zoveel mogelijk fysieke en logische standaarden toe te passen worden afhankelijkheden gespreid en een aantal risico's vermeden.

Naast deze specifieke aandachtspunten in functie van duurzame digitale archivering gelden natuurlijk ook de algemene vereisten voor een opslagsysteem. Factoren zoals beschikbaarheid (enkel tijdens kantooruren? 24/7?), performantie, snelheid, responsetijden, beveiliging, enz. dienen in overweging genomen te worden bij de keuze van een bepaald opslagsysteem.

Algemene vuistregels voor een duurzame en veilige digitale opslag zijn:

- zorg voor een welbepaalde beheersprocedure in functie van de gebruikte opslagmedia en de mogelijke risico's
- kies een opslagtechnologie die zijn betrouwbaarheid al bewezen heeft en waarover al expertise en ervaring beschikbaar is
- registreer bij archivering de CRC of de checksum van het digitaal object
- herbereken de CRC's of de checksum en vergelijk met de oorspronkelijke waarde. Doe dit continu en systematisch. Indien dit niet mogelijk is, doe dit met een vaste periodiciteit of steekproefgewijs met een representatief staal.

- spreid het risico: bewaar meerdere kopieën van hetzelfde digitaal object op verschillende types opslagmedia, of op opslagmedia van verschillende producenten, bijv.:
 - bewaar moederkopie op een tape, en reservekopie op een optische drager
 - gebruik voor de moederkopie blanco CD's van producent X en voor de reservekopie blanco CD's van producent Y
- documenteer het gebruik van opslagmedia. Houd documentatie bij over de indelingswijze en formattering van de opslagmedia.
- zorg voor herstelmogelijkheden: maak reservekopieën of back-ups van het digitaal archief. Vermijd dat er reservekopieën/back-ups van corrupte digitale objecten worden gemaakt. Controleer eerst de bitintegriteit van de digitale objecten alvorens er een reservekopie/back-up wordt gemaakt.
- houd de compatibiliteit van de gegevensdragers met de nieuwe generaties leesapparaten in de gaten. Zet de digitale objecten naar een nieuwe gegevensdrager over van zodra de nieuwe leesapparaten niet meer compatibel zijn met oude generaties van dezelfde types gegevensdrager.
- pas normen en standaarden toe:
 - gebruik enkel gestandaardiseerde types opslagmedia die door meerdere leveranciers worden ondersteund: gebruik geen opslagmedia die afhankelijk zijn van één leverancier of één bepaald systeem
 - bewaar de digitale objecten in een genormeerd bestandssysteem
- voorzie een afdoende beveiliging voor het opslagsysteem:
 - in geval van opslag in huis: afgesloten ruimte, toegangscontrole, stroomaggregatoren en koeling, enz.
 - in geval van opslag buitenshuis: dedicated verbindingen, gebruik van encryptie bij het verzenden van digitale objecten

4. DIGITALE DOCUMENTEN

De volgende stap in het archiveringsproces is een oplossing uitwerken om de leesbaarheid van digitale informatie op lange termijn te verzekeren. De moeilijkheidsgraad hangt in ruime mate af van de complexiteit en het bestandsformaat van het digitaal document. Digitale documenten worden immers bewaard in een specifiek formaat en zijn pas leesbaar wanneer de nodige applicatie- of viewsoftware voorhanden is. In bepaalde gevallen is zelfs een specifieke versie van de applicatiesoftware vereist. Software is echter het onderwerp van technologische veroudering zodat hierop tijdig dient te worden geanticipeerd.

Voor audiovisuele archieven volstaat het niet om alleen aandacht te besteden aan het bestandsformaat. Voor raadpleging van digitaal geluid of digitale video dienen ook de toegepaste codecs beschikbaar te zijn. De meeste formaten voor audiovisuele documenten zijn immers zogenaamde 'wrapperformaten' waarbinnen verschillende codecs voor geluid en bewegend beeld kunnen worden gebruikt. Of de audiovisuele documenten leesbaar zijn, hangt bijgevolg af van de beschikbaarheid van de overeenstemmende decoders.

De vraag hoe digitale documenten in een bepaald bestandsformaat op termijn raadpleegbaar blijven, houdt de IT- en de archiefwereld al jarenlang bezig. Oplossingen worden gezocht aan zowel de zijde van de digitale objecten, als aan de zijde van de technologie. Bij migratie worden originele bestandsformaten en de codecs omgezet naar meer duurzame bestandsformaten en codecs zodat ze met eigentijdse technologieën raadpleegbaar blijven. Een migratievoorbeeld is de omzetting van MS Word-documenten naar een preserveringsformaat zoals ODF of PDF/A. De emulatiestrategie gaat uit van de bewaring van originele bestandsformaten en codecs en bouwt de vereiste technologie na op een ander platform. Een emulatietoepassing is een 64-bits laptop met Windows Vista als besturingssysteem die zich gedraagt als een 16-bits x86-pc met DOS als besturingssysteem. Andere

preserveringsstrategieën zoals het omzetten van applicatieformaten naar eigentijdse versies (conversie), het bewaren van de originele hard-en software of digitale archeologie worden als niet realistisch beschouwd of lijken enkel een oplossing op korte termijn te bieden.

Een definitieve oplossing voor het leesbaarheidsprobleem is er nog niet. Wel is het duidelijk dat migratie van de documenten naar een geschikt archiveringsformaat en emulatie van de vereiste hard-en/of softwareomgeving als mogelijke oplossing elkaar niet uitsluiten, maar veeleer complementair zijn¹¹. Beide strategieën zijn geschikt voor welbepaalde type digitale documenten zodat het goed mogelijk is dat eenzelfde archiefbeherende instelling beide strategieën toepast. Migratie en emulatie kunnen ook binnen de levenscyclus van hetzelfde document worden gehanteerd.

De meeste opties blijven open wanneer het digitaal document in zowel zijn oorspronkelijk als zijn geschikt archiveringsformaat wordt bewaard. De DAVID-strategie om de leesbaarheid te bewaren, gaat hiervan uit¹². Deze strategie houdt in dat ten laatste op het archiveringsmoment het document naar zijn geschikt archiveringsformaat wordt omgezet, en dat zowel de oorspronkelijke als de gemigreerde representatie in het archiveringssysteem worden opgenomen. Op die manier blijven de volgende opties voor raadpleging in de toekomst mogelijk:

- emulatie van het document in zijn oorspronkelijk formaat
- emulatie van het document in zijn geschikt archiveringsformaat
- migratie van het document op basis van het oorspronkelijk formaat
- migratie van het document op basis van zijn geschikt archiveringsformaat

Het bewaren van digitale documenten in een geschikt archiveringsformaat speelt een essentiële rol in het bewaren van de leesbaarheid. In de keuze van een geschikt archiveringsformaat geldt eveneens de stelregel dat afhankelijkheid ten aanzien van een specifieke softwarepakket absoluut te vermijden is. In die zin is het bewaren van digitale documenten in producentgebonden en niet-gedocumenteerde formaten zoals die van MS Office¹³ of AutoCAD heel risicovol. Beter is om een digitaal document in een formaat te bewaren dat door meerdere applicaties wordt ondersteund. Hetzelfde geldt voor het kiezen van archiveringscodecs voor audiovisuele documenten. Bij voorkeur worden gestandaardiseerde en geen producentafhankelijke codecs gebruikt. Een tweede aandachtspunt heeft betrekking op het al dan niet toepassen van compressie. Het gebruik van compressie bij dataopslag heeft als voordeel dat er wordt bespaard op vereiste opslagcapaciteit, bandbreedte bij transmissie en dat gecomprimeerde bestanden sneller worden verwerkt of gepresenteerd. Het nadeel van compressie is dat extra schakels of afhankelijkheden aan het reconstructieproces worden toegevoegd en dat in het geval van lossy compressie kwaliteits- en informatieverlies optreedt. In volgorde van voorkeur kan men kiezen voor opslag zonder compressie, opslag met lossless compressie of opslag met lossy compressie¹⁴. Voor digitale afbeeldingen en digitaal geluid is compressieloze opslag technisch haalbaar, al hangt daar natuurlijk een prijskaartje aan vast. Bij digitaal bewegend beeldmateriaal is compressieloze opslag minder evident. Bewegend beeld digitaal archiveren vergt een heel grote opslagcapaciteit zodat compressie nagenoeg onvermijdbaar is. In dit geval is het aangewezen om een genormeerde compressiemethode toe te passen.

Normering is een belangrijke kwaliteitsvereiste van een geschikt archiveringsformaat of geschikte archiveringscodec, maar is lang niet het enige criterium bij de keuze van een geschikt

¹¹ Voor een uitgebreide evaluatie van de beschikbare digitale bewaarstrategieën, zie: F. BOUDREZ, *B. Digitale bewaarstrategieën*, in: F. BOUDREZ en H. DEKEYSER, *Digitaal archiveren in de praktijk. Handboek*, Antwerpen, 2004. (beschikbaar op: <http://www.edavid.be/davidhandboek>).

¹² F. BOUDREZ, *B. DIGITALE BEWAARSTRATEGIEËN*, in: F. BOUDREZ en H. DEKEYSER, *Digitaal archiveren in de praktijk. Handboek*, Antwerpen, 2004. (beschikbaar op: <http://www.edavid.be/davidhandboek>)

¹³ De binaire formaten van MS Office tot en met versie 2003 zijn niet volledig gedocumenteerd. MS Office 2007 gebruikt een nieuw bestandsformaat dat op XML is gebaseerd. De specificatie van het MS Office 2007 formaat is wel vrijgegeven.

¹⁴ Bij *lossless* compressie gaat geen informatie of kwaliteit verloren. Het gedecomprimeerde digitaal object is identiek aan het oorspronkelijke digitaal object. Dit in tegenstelling tot *lossy* compressie, waarbij het gedecomprimeerde digitaal object niet identiek is aan het oorspronkelijke digitaal object.

archiveringsformaat. De correcte en ongewijzigde overname van de essentiële eigenschappen van het document in het nieuwe formaat is minstens even belangrijk. In het geval van digitale archiefdocumenten worden die essentiële eigenschappen hoofdzakelijk gedefinieerd vanuit de vereisten die de werkprocessen stellen ten aanzien van de archiefdocumenten. Bij omzettingen dienen minimaal die essentiële eigenschappen ongewijzigd te worden omgezet. Dit bepaalt mee welk formaat als archiveringsformaat wordt gekozen en welke omzettingstool wordt gebruikt.

Er zijn voor verschillende documenttypes meerdere bestandsformaten beschikbaar die als archiveringsformaat kunnen worden gebruikt. Voor tekstverwerkingsdocumenten, spreadsheets en presentaties bijvoorbeeld zijn ODF¹⁵ en in mindere mate PDF/A¹⁶ bruikbaar als archiveringsformaat, voor e-mail en databases is dit XML¹⁷. Voor rasterafbeeldingen worden TIFF en JPEG(2000) veel gebruikt. Voor de documenttypes waarvoor geen genormeerde formaten beschikbaar zijn, gebruikt men best een uitwisselingsformaat als archiveringsformaat (bijv. DXF voor CAD-tekeningen). Uitwisselingsformaten zijn doorgaans ook goed gedocumenteerd en zijn met meerdere applicaties leesbaar. In de praktijk zullen de meeste archiefbeherende instellingen het aantal archiveringsformaten die ze hanteren, proberen in de hand te houden en op dit vlak een aantal 'standaarden' hanteren. Voor elk archiveringsformaat is immers op termijn een conserveringsstrategie nodig.

Samen met het vastleggen van de 'standaarden' op het vlak van archiveringsformaten is de keuze van een conserveringsstrategie een belangrijk onderdeel van het bepalen van het digitaal archiveringsbeleid. Tot op heden blijft migratie de meest toegepaste strategie. Migratie wordt algemeen beschouwd als meer praktisch haalbaar: er zijn veel migratietools voorhanden, gemigreerde documenten zijn direct raadpleegbaar, de archiefbeheerder hoeft enkel documenten en geen software te archiveren, archiefgebruikers werken met eigentijdse software, vergelijkingen met het document in zijn oorspronkelijk formaat blijven mogelijk, enz. Digitale documenten omzetten van het ene bestandsformaat naar het andere zijn echter operaties met een grote impact op de digitale informatie. Deze operaties worden best zorgvuldig gepland en gecontroleerd zodat er geen noemenswaardig of significant kwaliteitsverlies optreedt. Net zoals bij de bewaring van digitale objecten is de omzetting naar een geschikt archiveringsformaat evenmin een permanente oplossing. Ook normen zijn onderhevig aan technologische veroudering zodat nieuwe omzettingen zich zullen opdringen van zodra hun ondersteuning dreigt te verdwijnen.

Dit probleem stelt zich overigens niet alleen voor migratie. Ook emulatieplatformen worden op termijn geconfronteerd met technologische veroudering. Op dat moment dienen ofwel nieuwe emulatieprogramma's te worden gemaakt of dient een emulator voor de emulator te worden gemaakt. Wil men de emulatie optie op termijn open houden, dan volstaat het evenmin om de digitale documenten in hun oorspronkelijk formaat te bewaren. De huidige emulatiemethode voor digitale archivering richt zich op de computerhardware, wat betekent dat de originele besturingssystemen en applicatiesoftware dienen te worden gearchiveerd¹⁸. Om emulatie als optie mogelijk te houden, zal het dus niet volstaan om enkel de digitale documenten in hun oorspronkelijk formaat te archiveren.

Bij emulatie spelen ook nog een viertal andere kwesties. Ten eerste vraagt het maken van emulatieplatformen veel middelen en heel gespecialiseerde kennis. De vraag is niet alleen welke organisaties of instellingen de nodige kennis hebben en/of deze investering kunnen dragen, maar ook of die inspanning op termijn vol te houden is. Ten tweede gaat de bestaande emulatiemethode voor digitale preservatie uit van de archivering van de oorspronkelijke besturingssystemen en

¹⁵ ISO/IEC 26300(2006): Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0.

¹⁶ ISO 19005-1(2005): Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1).

¹⁷ Extensible Markup Language (XML) 1.0 (Fourth Edition): W3C Recommendation 16 August 2006. Een XML Schema voor de archivering van e-mails als XML-documenten is beschikbaar op: <http://www.expertisecentrumdavid.be/xmlschemas/email.xsd>.

¹⁸ In de praktijk wordt hiervoor niet de installatiesoftware bewaard, maar worden geïnstalleerde versies in de vorm van ghostimages of disk images gearchiveerd.

applicatiesoftware. De vraag is hier of dit niet strijdig is met de auteurswetgeving. Wie software 'koopt', betaalt een gebruiksrecht, maar wordt geen eigenaar van de software en mag die in principe niet verder verspreiden of beschikbaar stellen voor derden. Ten derde houdt emulatie in dat de eindgebruiker met oude (applicatie-)software moet kunnen werken. Wie kan binnen 10 jaar nog werken met DOS 6.22, WordPerfect 5.1 of Lotus 1-2-3? Ten slotte staat de ervaring met emulatie voor langetermijnarchivering nog in zijn kinderschoenen. Emulatie als strategie wordt al langer toegepast binnen andere vakdisciplines, maar momenteel is er bijvoorbeeld nog maar 1 emulator gemaakt voor langetermijnarchivering (de Dioscuri¹⁹).

Het bewaren van de leesbaarheid van digitale documenten op (middel-)lange termijn vraagt een conserveringsstrategie. Hiervoor dienen een aantal keuzes te worden gemaakt, zoals: welke strategie wordt toegepast, welke tools worden hierbij gebruikt, met welke instellingen worden die tools gehanteerd, enz. Om deze beslissingen mee aan te sturen en onmiddellijk ook te documenteren kan PLATO²⁰ worden gebruikt. Dit is een (online) planningstool waarbij 4 stappen worden doorlopen: het bepalen van de vereisten, het onderzoeken van alternatieven, het analyseren van de resultaten en het definiëren van de conserveringsstrategie.

Algemene vuistregels voor een duurzame leesbaarheid zijn:

- archiveer digitale documenten niet uitsluitend in een producent- of applicatieafhankelijk formaat, maar bewaar ze ook in een geschikt archiveringsformaat. Een geschikt archiveringsformaat is o.a.:
 - genormeerd of minimaal een industriestandaard met de status van uitwisselingsformaat
 - goed gedocumenteerd
 - wijdverspreid en breed ondersteund
 - in staat om de essentiële eigenschappen van het document ongewijzigd in de tijd over te brengen.
- neem het document in zijn oorspronkelijk en zijn geschikt conserveringsformaat in het archiveringssysteem op
- plan zorgvuldig het migratiepad:
 - werk een transparant en validerend migratieproces uit
 - test de migratie op voorhand heel zorgvuldig: zorg voor een representatieve steekproef
 - documenteer de migratie operaties: beschrijf het migratieproces (bronformaat, doelformaat, migratiesoftware, instellingen van de migratiesoftware, eigenschappen van documenten die wijzigen) en registreer welke digitale objecten werden omgezet
- documenteer de relatie tussen:
 - de documenten en hun verschijningsvormen
 - de verschijningsvormen en hun digitale objecten
- registreer de technische kenmerken van de digitale objecten die je in het digitaal archief opneemt: bestandsformaat, versienummer, formaatprofiel, encoding, XML namespace, enz.²¹
- vermijd compressie. Indien compressie onvermijdbaar is (bijv. bij het archiveren van bewegend beelden) kies dan in de mate van het mogelijke voor een lossless compressiemethode. Zorg er in ieder geval voor dat de (de-)compressiemethode open en gedocumenteerd is.
- volg de technologische evolutie op en grijp in van zodra de ondersteuning van het bestandsformaat en/of de codec dreigt weg te vallen.

¹⁹ <http://dioscuri.sourceforge.net/>

²⁰ PLATO: Planets preservation planning tool. PLATO is online beschikbaar op: <http://www.ifs.tuwien.ac.at/dp>

²¹ Open source tools die hiervoor kunnen worden gebruikt, zijn o.a. JHOVE en DROID.

5. DIGITALE ARCHIEFDOCUMENTEN

Voor de archivering van begrijpbare en betrouwbare digitale archiefdocumenten volstaat het niet dat ze leesbaar blijven. Vanwege hun archiefstatus gelden nog bijzondere kwaliteitseisen. Archiefdocumenten moeten vindbaar, bruikbaar, authentiek, integer en betrouwbaar zijn. Om dit te realiseren zijn metadata nodig. Met behulp van metadata wordt de herkomst en het beheer gedocumenteerd, de 'significant properties' of essentiële eigenschappen van de documenten vastgelegd, de betrouwbaarheid onderbouwd, de toegankelijkheid vergroot, enz. De metadata van digitaal archief zijn in de regel ook in digitale vorm geregistreerd. Bijgevolg gelden dezelfde duurzaamheidsvereisten voor de metadata als voor de digitale documenten waarop ze betrekking hebben.

De metadata van digitale archiefdocumenten worden doorgaans in de specifieke bedrijfsapplicaties van de archiefbeherende instellingen opgeslagen. Dit betekent echter niet automatisch dat de metadata van of over de digitale archiefdocumenten vrij zijn van risico's. Digitale archiefdocumenten en hun metadata worden doorgaans gescheiden van elkaar opgeslagen en/of er wordt een specifieke technologie voor het koppelen gebruikt. Hierdoor bestaat het risico dat hun relatie verbroken geraakt. Dit verlies heeft een grote impact, want hierdoor verliest het digitaal archiefdocument de facto zijn archiefstatus.

Idealiter is er een onlosmakelijke band tussen de digitale archiefdocumenten en hun metadata. Dit kan bijvoorbeeld gerealiseerd worden door inkapseling of inbedding toe te passen. Beide benaderingen hebben met elkaar gemeen dat het digitaal archiefdocument en hun metadata, die een logisch geheel vormen, ook fysiek samen worden gearhiveerd en dat de metadata rechtstreeks in het digitaal archief wordt bewaard. Opteert men niet voor inkapseling of inbedding, dan verschuift de aandacht en de inspanningen naar het controleren en herstellen van de band tussen het digitaal archiefdocument en zijn metadata. Dit wordt dan een heel belangrijke beheerstaak.

Bij inbedding worden metadata aan de digitale documenten toegevoegd. De meeste bestandsformaten voorzien een aantal metadatavelden in hun bestandsheader waarvan gebruik kan worden gemaakt. Hierbij heeft men de keuze tussen gebruik maken van de verschillende metadatavelden die een bepaald bestandsformaat voorziet of alle metadata als XML in één metadataveld te stoppen²². In deze laatste optie kan men vrij zijn metadatavelden en metadatastructuur kiezen. Inbedding heeft echter een aantal nadelen, waardoor deze methode niet zo geschikt is:

- voor elk bestandsformaat is ondersteuning nodig om metadata toe te voegen of uit te lezen
- bij nieuwe migratie-operaties moet de ingebedde metadata mee worden overgezet
- inbedding in bestandsheaders leidt tot reusachtige redundantie: contextuele metadata wordt dan niet altijd op het niveau van de logische entiteit 'het archiefdocument' bijgehouden, maar op het niveau van de digitale objecten. Er is immers niet altijd een één-op-één relatie tussen digitale archiefdocumenten en digitale objecten.

Deze nadelen kunnen met inkapseling allemaal worden vermeden. Bij inkapseling worden de metadata en het digitaal archiefdocument verpakt in één computerbestand. Door één inkapselingsformaat te hanteren, dient men slechts voor één bestandsformaat de mogelijkheid voorzien om metadata toe te voegen of uit te lezen. Wanneer de verschillende representaties van hetzelfde archiefdocument en hun digitale objecten in één fysiek containerbestand worden ingekapseld, dan hoeven de contextuele en beschrijvende metadata slechts één keer te worden opgeslagen. Het risico op verlies van metadata bij nieuwe omzettingen wordt vermeden wanneer de nieuwe representatie aan het bestaande containerbestand wordt toegevoegd.

Als formaat voor die ingekapselde computerbestanden heeft men de keuze uit bijvoorbeeld PDF Packages, ZIP, JAR en natuurlijk ook XML. Vooral XML is een heel interessante keuze, want op die

²² Deze laatste optie werd onderzocht en uitgewerkt door S. HEUSCHER, *Persistent and integer lineage for digital objects. The SIMPLE approach*, Bern, 2006.

manier kan men XML combineren als inkapselings-, metadata- en preserveringsformaat²³. Door XML te kiezen, heeft de archiefbeheerder ook de vrijheid om zijn eigen XML Schema te hanteren. Hierdoor ontstaat de mogelijkheid om een XML structuur in functie van het eigen metadatamodel en de eigen informatiearchitectuur te ontwerpen. Een tweede voordeel is dat met XML de duurzaamheid van het inkapselingsformaat maximaal is verzekerd. Dit zijn belangrijke elementen wil men het digitaal archief volledig onder controle hebben.

Het toepassen van inkapseling of inbedding betekent echter niet dat geen metadata meer gecentraliseerd in archiefbeherende systemen en hun databases worden opgeslagen. Om performante en transparante zoekfunctionaliteiten mogelijk te maken, is centralisering van metadata in databasesystemen aangewezen. Bovendien beheren archiefbeherende instellingen ook niet-digitale archieven waarvoor eveneens metadata aanwezig zijn. De ingekapselde of ingebedde metadata kunnen dan beschouwd worden als reservekopieën en/of kunnen beperkt worden tot de echt essentiële contextuele en beschrijvende metadata.

Opslag van metadata in archiefbeherende systemen betekent ook niet automatisch dat de metadata op een digitaal duurzame wijze of zelfs op een transparante wijze zijn opgeslagen. De databases van de archiefbeherende systemen staan veelal buiten het digitale archiveringssysteem (bijv. het digitale depot) of zijn maar zelden het voorwerp van een digitaal archiveringsbeleid. Archivarissen ontwerpen archiveringsstrategieën voor tal van bedrijfsapplicaties van de archiefvormers waarvoor ze bevoegd zijn, maar is er ook een archiveringsstrategie voor hun eigen archiefbeheerssysteem? De situatie wordt nog problematischer wanneer blijkt dat hun eigen archiefbeheerssysteem weinig open en/of amper gedocumenteerd is. In bepaalde gevallen kan dit zelfs leiden tot een heuse vendor lock-in voor de metadata, of zelfs voor het volledige digitale archiveringssysteem.

Algemene vuistregels voor duurzame archiefdocumenten zijn:

- archiveer de metadata van of over de digitale archiefdocumenten op een gestructureerde en digitaal duurzame wijze
- zorg voor een duurzame band tussen de digitale archiefdocumenten en hun metadata:
 - pas inkapseling of inbedding toe
 - controleer de relatie tussen de digitale archiefdocumenten en hun metadata
- maak het metadata- en het databasemodel van het archiefbeheerssysteem je eigen
- zorg voor een afdoende beveiliging bij de opslag, het beheer en de raadpleging van de digitale archiefdocumenten
- leg vast wat de betekenisgevende of essentiële eigenschappen van het 'origineel' of 'authentiek' document zijn. Zorg ervoor dat deze eigenschappen ongewijzigd in tijd worden overgebracht zodat de documenten hun originele functie of betekenis behouden.
- ontwikkel een archiveringsstrategie voor het archiefbeheerssysteem.

6. BESLUIT

Door zijn aard verschilt digitale archivering op een aantal punten fundamenteel met de archivering van bijvoorbeeld papieren informatie. Preservering van papieren documenten is mogelijk mits een goede conservering van de gegevensdragers en heeft de bewaring van de originele documenten in ongewijzigde vorm als uitgangspunt. De archivering van digitale informatie is in veel gevallen enkel mogelijk mits vervanging van de gegevensdragers, conversies, migraties, verrijking met metadata, enz. Zonder deze ingrepen is digitale informatie gedoemd om te verdwijnen.

Duurzame digitale archivering vraagt een heel actief preserveringsbeleid dat zich richt op de intacte en compatibele opslag van digitale objecten, de leesbaarheid van bestandsformaten en gebruikte codecs,

²³ F. BOUDREZ, *Digitale containers voor het archief*, Antwerpen, 2005. De XML Schemas die voor deze strategie door eDAVID werden ontworpen, zijn beschikbaar op: www.edavid.be/xmlschemas

en op de archivering van metadata in relatie tot de documenten waarop ze betrekking hebben. Voor elk aspect van digitale archivering dienen passende oplossingen worden gezocht. Digitale archivering als een risicovol reconstructieproces beschouwen, levert hiervoor het kader en de basisprincipes. Geen enkele oplossing of aanpak is echter definitief of permanent. Door de blijvende afhankelijkheid van hard- en software in het algemeen, zal men blijvend moeten anticiperen op de technologische evoluties. Idealiter wordt hiermee niet gestart bij opname in een archiveringssysteem, maar omspannt het digitale archiveringsbeleid de volledige levenscyclus van een archiefdocument.