

Digitale documenten archiveren. Aandachtspunten en vuistregels

Filip Boudrez
Expertisecentrum DAVID vzw
Antwerpen, 2007

0. INHOUD

1. Inleiding.....	1
2. Digitaal archiveren als reconstructieproces.....	1
3. Bewaren van digitale objecten.....	2
4. Bewaren van digitale documenten.....	3
5. Bewaren van digitale archiefdocumenten.....	5
6. Besluit.....	6

1. INLEIDING

Nagenoeg iedereen die vandaag de dag bij de uitoefening van zijn taken en activiteiten informatie nodig heeft, doet dat zoveel mogelijk digitaal. Digitale informatie is immers gemakkelijk herbruikbaar en snel uitwisselbaar. Het lijkt dan ook evident dat die informatie op digitale wijze wordt bewaard en in de toekomst raadpleegbaar blijft. Nochtans is dit niet zo vanzelfsprekend als het misschien wel lijkt. Digitale archivering houdt immers een aantal aandachtspunten in, die we hier kort willen toelichten.

2. DIGITAAL ARCHIVEREN ALS RECONSTRUCTIEPROCES

Het archiveren van digitale informatie verschilt in een aantal opzichten fundamenteel met het bewaren van papieren documenten. Veel van die verschillen vinden hun oorsprong in WAT in de papieren en de digitale wereld eigenlijk wordt opgeslagen. In de papieren wereld vormen de gegevensdrager en het document een fysieke eenheid. Hierdoor geldt voor papieren documenten dat wat je archiveert en achteraf raadpleegt identiek hetzelfde is, nl. het document. In de digitale wereld is dit niet zo. Van een digitaal document bewaren we de bits en bytes. Die bits en bytes vormen geen fysieke eenheid met hun opslagmedium, en zijn evenmin hetzelfde als het document. Wat je in de digitale wereld raadpleegt, is een document dat telkens opnieuw op basis van de opgeslagen bits en bytes moet worden opgebouwd. Het digitaal document moet bij raadpleging als het ware telkens opnieuw worden gereconstrueerd¹.

Digitale archivering als een reconstructieproces beschouwen, biedt een goed perspectief om de noden van efficiënte digitale archivering in kaart te brengen. 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

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).

- digitale archiefdocumenten: de gearchiveerde documenten zijn bruikbaar, wat impliceert dat ze begrijpbaar en betrouwbaar zijn.

Koppelt men deze benadering van digitale archivering aan risk assessment, dan kunnen onmiddellijk al twee belangrijke conclusies worden geformuleerd. Ten eerste, uit hoe meer schakels het reconstructieproces bestaat, des te kwetsbaarder of risicovol het proces is. Van zodra één schakel in het reconstructieproces ontbreekt, dient het digitaal archiefdocument als verloren beschouwd te worden. 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 zijn het gebruik van compressie, encryptie, paswoorden, enz. beter te vermijden. Niet alle reconstructieschakels kunnen echter worden vermeden. Finaal zullen nog steeds hard- en software nodig zijn voor de raadpleging. De tweede conclusie is dan ook dat voor de overblijvende schakels de informatiebeheerder 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 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.

3. BEWAREN VAN DIGITALE OBJECTEN

Elke toepassing van digitale archivering vraagt minimaal dat computerbestanden op een duurzame wijze in tijd kunnen worden overgebracht. Digitale opslag evolueert echter heel snel. Enerzijds is er de evolutie naar een steeds grotere opslagcapaciteit, waardoor ook de impact van een defecte gegevensdrager almaar toeneemt. Anderzijds zoekt de industrie verder naar robuustere en duurzame opslagoplossingen, waarbij voortdurend tussen magnetische, optische of nieuwe vormen van opslagmedia wordt gependeld². Digitale gegevensdragers blijven immers onderhevig aan natuurlijke degeneratie en gebruiksslijtage.

Elke informatiebeheerder zoekt duurzame gegevensdragers voor de opslag van zijn digitale objecten. In zijn zoektocht wordt hij/zij bestookt met enerzijds langetermijngaranties van producenten en met anderzijds onheilspellende berichten over de duurzaamheid van CD's, onleesbare DVD's, gecrashte harde schijven, gesprongen tapes en andere calamiteiten. Een goed uitgangspunt hierbij is de veronderstelling dat de levensduur van de vereiste afspeeltechnologie korter is dan de levensduur van een duurzame gegevensdrager. In die zin maakt het weinig uit of een CD een levensduur van 30 jaar dan wel van 100 jaar heeft. De vraag is over hoe lang we over de vereiste technologie zullen beschikken om een CD te lezen. Om van een zo groot mogelijke ondersteuning te genieten is het belangrijk om gestandaardiseerde types gegevensdrager te kiezen³.

Een duurzame gegevensdrager is overigens niet alleen een kwestie van een type gegevensdrager te kiezen die van nature uit niet of slechts heel traag degradeert. De duurzaamheid is ook afhankelijk van de wijze waarop de gegevens op het opslagmedium zijn geordend of gestructureerd. Met name het bestandssysteem ('formatting') is hierbij van belang. Het afspeelapparaat en/of de computer moeten de indelingswijze van de gegevens op de drager ondersteunen. Niet voor elk type gegevensdrager kan echter een genormeerd bestandssysteem worden toegepast. Dit is bijvoorbeeld het geval voor harde schijven en voor DVD's. De eerste generatie DVD had hier al onder te leiden (DVD-R, DVD-RW, DVD+R, DVD+RW), en momenteel woedt de formaat oorlog bij de tweede generatie DVD (Blu-Ray vs HD

² Bijvoorbeeld: holografische schijven, DVD's met een glazen substraatlaag, enz.

³ Via de website www.iso.ch kan gemakkelijk worden geverifieerd of een bepaald type gegevensdrager en/of bestandssysteem de status van ISO-norm heeft. Indien ja, dan wordt best nagegaan in welke mate er ondersteuning is door verschillende technologieën of producenten.

DVD) volop. Ondertussen is de derde generatie DVD al aangekondigd voor 2010. Deze optische schijven zouden een opslagcapaciteit tot 1 TB hebben.

Het toepassen van fysieke en logische standaarden biedt garanties voor de leesbaarheid van de gearchiveerde bits en bytes, maar garandeert nog niet dat de computerbestanden intact zijn. Bits en bytes dreigen wel eens omver te vallen. Om dit te vermijden of achteraf te kunnen herstellen zijn bijkomende maatregelen nodig. Gegevensverlies kan vermeden worden door reservekopieën te maken, redundante opslagsystemen te gebruiken en kwaliteitscontroles uit te voeren.

Vuistregels:

- gebruik een genormeerd type gegevensdrager (bijv. DLT of LTO voor magnetische tapes, CD voor optische dragers).
- hanteer een genormeerd bestandssysteem (bijv. ISO-9660 voor CD-ROM)
- kies een gegevensdrager die voorziet in een robuust foutopsporings- en foutverbeteringsmechanisme
- spreid het risico door:
 - veiligheidskopieën (off-site) van de gegevensdragers bij te houden:
 - hoe groter de opslagcapaciteit, hoe meer veiligheidskopieën nodig zijn
 - eventueel een ander type gegevensdrager voor de veiligheidskopieën te gebruiken
 - bij opslag op harde schijven:
 - een opslagsysteem met redundantie of pariteitsinformatie te gebruiken
 - een mirror in een ander bestandssysteem bij te houden
 - back-ups aan te maken
- bewaar de gegevensdragers in optimale omstandigheden
- controleer regelmatig, of indien mogelijk systematisch, de bitintegriteit van de digitale objecten (bijv. door checksums te herberekenen en te vergelijken)
- documenteer het gebruik en het beheer van de gegevensdragers.

Meer informatie over de duurzame opslag van digitale objecten op magnetische en optische gegevensdragers is beschikbaar op de website van eDAVID⁴.

4. BEWAREN VAN DIGITALE DOCUMENTEN

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

De vraag hoe digitale documenten in een bepaald bestandsformaat op termijn raadpleegbaar blijven, houdt de IT- en de archiefwereld al jarenlang bezig. Een definitieve oplossing 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 zijn⁵. De meeste opties blijven beschikbaar wanneer het digitaal document in zowel zijn oorspronkelijk als zijn geschikt archiveringsformaat wordt bewaard. De DAVID-strategie om de leesbaarheid te bewaren, gaat hiervan uit.

⁴ F. BOUDREZ, *CD's voor het archief*, Antwerpen, 2001; Digitaal Archiveren: richtlijn & advies, nr. 2: *Duurzame CD's*; F. BOUDREZ, *Magnetische dragers voor het archief*, Antwerpen, 2002; Digitaal Archiveren: richtlijn & advies, nr. 6: *Duurzame magnetische dragers*. (www.edavid.be/davidproject).

⁵ 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>)

Ook hier geldt 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 Office⁶ of AutoCAD heel risicovol. Beter is om een digitaal document in een formaat te bewaren dat door meerdere applicaties wordt ondersteund. Normering biedt hier opnieuw enkele garanties voor, maar is lang niet het enige criterium bij de keuze van een geschikt archiveringsformaat. De correcte en ongewijzigde overname van de essentiële eigenschappen van het document is minstens even belangrijk. Voor tekstverwerkingsdocumenten, spreadsheets en presentaties is ODF⁷ en in mindere mate PDF/A⁸ een geschikt archiveringsformaat, voor e-mail en databases is dit XML⁹.

Digitale documenten omzetten van het ene bestandsformaat naar het andere zijn 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.

Vuistregels:

- archiveer digitale documenten niet uitsluitend in een producent- of applicatieafhankelijk formaat, maar bewaar ze ook in een geschikt archiveringsformaat. Documenteer de omzettingen.
- vermijd compressie. Indien compressie onvermijdbaar is (bijv. bij het archiveren van bewegend beelden) kies dan in de mate van het mogelijke voor een lossless compressiemethode¹⁰. Zorg er in ieder geval voor dat de (de-)compressiemethode open en gedocumenteerd is.
- kies een archiveringsformaat in functie van:
 - de normering en beschikbare ondersteuning
 - de mate waarin de essentiële eigenschappen van het digitaal document worden behouden
- werk een transparant, gedocumenteerd en validerend proces uit voor de omzetting naar andere bestandsformaten
- documenteer essentiële eigenschappen die dreigen te verdwijnen of te wijzigen als gevolg van omzettingen in de vorm van metadata
- leg de nodige technische gegevens (bijv. bestandsformaat, formaatprofiel, encoding, enz.) vast die nodig zijn om de leesbaarheid op lange termijn te verzekeren
- volg de technologische evolutie op

⁶ De binaire formaten van MS Office tot en met versie 2003 zijn niet gedocumenteerd. MS Office 2007 gebruikt een nieuw bestandsformaat dat op XML is gebaseerd. De specificatie van het MS Office 2007 formaat is wel vrijgegeven, in een poging om dit formaat als een ISO-norm te laten vastleggen. Dit laatste is vooralsnog niet gelukt.

⁷ ISO/IEC 26300(2006): Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0

⁸ ISO 19005-1(2005): Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1).

⁹ 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>.

¹⁰ Bij *lossless* compressie gaat geen informatie of kwaliteit verloren. Dit in tegenstelling tot *lossy* compressie.

5. BEWAREN VAN DIGITALE ARCHIEFDOCUMENTEN

Digitale documenten zijn pas begrijpbaar wanneer bekend is binnen welke context ze werden gecreëerd of ontvangen. In enkele gevallen bevatten de documenten een verwijzing naar die context, maar doorgaans is die informatie niet expliciet aanwezig. Voor een optimaal hergebruik is het belangrijk dat de relatie van een document met een dossier of een onderwerp op een expliciete wijze wordt aangegeven zodat hun ontstaans- en gebruikscontext duidelijk is.

Eén mogelijkheid om dit te realiseren is de documenten voorzien van de nodige metadata zodat duidelijk wordt gemaakt binnen welke context ze werden opgemaakt, ontvangen en/of gebruikt. Deze aanpak vraagt om een grote mate van automatisering. Een andere en eenvoudiger manier is het vormen van dossiers of onderwerpsmappen. In de digitale mappen worden alle documenten die over die zaak of dat onderwerp handelen samen bewaard. E-mails en hun bijlagen worden best in de overeenstemmende dossier- of onderwerpsmap opgeslagen, en blijven niet gewoon in het e-mailsysteem staan. Binnen een Windows-desktopomgeving kun je bijvoorbeeld hiervoor in de Windows verkenner mappen aanmaken, die dan alle documenttypes bevatten die over dezelfde zaak of hetzelfde onderwerp handelen.

Door de documenten samen te klasseren, is het achteraf ook snel duidelijk welke informatie over een zaak of onderwerp aanwezig is en herleid je de mogelijke vindplaatsen van digitale informatie tot één locatie. De relatie van de dossiers- of onderwerpsmappen met de werkprocessen waarin ze tot stand kwamen, kan worden geëxpliciteerd door de digitale mappen te structureren en door hoofdmappen voor de werkprocessen te voorzien. Binnen die hoofdmappen maak je dan voor iedere concrete zaak een nieuwe subfolder met een betekenisvolle naam aan.

Naast contextualisering dient de informatiebeheerder er ook voor te zorgen dat de digitale archiefdocumenten betrouwbaar zijn, dat hun identiteit duidelijk is geregistreerd en dat hun essentiële componenten gefixeerd zijn¹¹. Deze archiefvereisten staan dikwijls in schril contrast met het 'digitaal-zijn'. Digitale documenten kunnen immers snel worden aangepast, bevatten zelden alle identificerende metadata en hun inhoud of opmaak wordt lang niet altijd op een gefixeerde wijze vastgelegd. Omwille van deze reden doen instellingen en ondernemingen een beroep op meer geavanceerde documentbeheerssystemen of records management applicaties voor het beheer van hun digitale documenten. Deze systemen zorgen er mee voor dat versiebeheer, toegangscontrole, (automatisch) registreren van metadata, enz. op de documenten wordt toegepast. Dezelfde aandachtspunten gelden ook voor databasetoepassingen. Pas wanneer hier rekening wordt gehouden bij de inrichting van informatiesystemen kunnen betrouwbare digitale archiefdocumenten worden gearchiveerd. Dit laatste illustreert nogmaals de nood van het proactief optreden bij digitaal documentbeheer en digitale archivering.

Vuistregels:

- bewaar alle digitale documenten met betrekking tot dezelfde zaak of hetzelfde onderwerp in één digitale map
- structureer de digitale mappen zodat hun relatie met de werkprocessen duidelijk wordt aangegeven
- klasseer in het dossier- of de onderwerpsmap eveneens de e-mails en hun bijlage(n)
- voorzie toegangscontrole tot de digitale documenten
- leg de essentiële componenten van het digitaal archiefdocument op een statische en onwijzigbare wijze vast

¹¹ Een voorbeeld van dit laatste is een automatisch datumveld in een tekstdocument. Telkens het document wordt geopend of afgedrukt, bevat dit veld de actuele datum.

6. BESLUIT

Digitale archivering vraagt een heel actieve vorm van preservering, waarbij voortdurend wordt geanticipeerd op de wijzigende technologische context. Bovendien betekent digitale archivering veel meer dan zorg dragen om de gegevensdrager alleen. Het leesbaar houden van digitale objecten en documenten vraagt een continu beheer. De risico's die inherent zijn aan het proces van digitaal archiveren vragen eveneens om extra voorzorgsmaatregelen en preserveringstaken.