

De bewaring van digitale informatie op lange termijn

1. Inleiding



Isabelle Boydens is consultant bij de sectie Onderzoek. Als Doctor in de Letteren en Wijsbegeerte, richting « Informatie- en Documentatiewetenschappen » geeft zij les aan de Franstalige Vrije Universiteit van Brussel (ULB). Haar werkzaamheden betreffen kwaliteitsanalyses van databases, meta-informatiesystemen (zoals de « elektronische glossaria » ontwikkeld in het kader van het project DMFA - ASR), documentaire toepassingen, indexerings- en zoekmethoden en collaboratieve systemen.

Contact : 02/509.59.92

isabelle.boydens@smals-mvm.be

1.1. Een eeuwenoude problematiek

De bewaring van digitale informatie op lange termijn is een cruciale kwestie geworden in de wereld van de ondernemingen en administraties. Wij stellen hier voor om dieper in te gaan op de factoren die aan de oorsprong liggen van het verval van de digitale informatie, alsook op de strategieën die dit moeten verhelpen. We zullen zien dat er bij deze kwestie talrijke factoren komen kijken zowel qua « hardware » en « software » als op conceptueel vlak¹.

Bij de aanvang van deze studie merken we op dat de problematiek niet nieuw is. Bij het begin van de Middeleeuwen maakte de overstap van papyrus naar perkament immers al een consolidatie van de fysieke dragers mogelijk². Maar het massale gebruik van dierenhuiden leidde tot economische moeilijkheden. Zo verschenen dan de palimpsesten, stukken perkament waarvan de oorspronkelijke geschriften verwijderd werden om er nieuwe tekst op neer te schrijven, net als bij de procedures die momenteel aangeboden worden door onze overschrijfbaar optische schijven. Een aantal geschriften werden dus gewist. Voor de uitvinding van de boekdrukkunst werden informatiesystemen uitgebouwd doorheen generaties van « monniken-kopiïsten » die de geschriften gedurende eeuwen overschreven. Deze geschriften werden ons slechts gedeeltelijk overgeleverd : het gebeurt dat het origineel verloren is gegaan en dat men slechts een onvolledige verzameling van uiteenlopende kopieën heeft (door fouten, al dan niet vrijwillig gemaakt door de kopiïsten : vervormde, weggelaten of toegevoegde passages). Om een tekst te reconstrueren en het origineel zo dicht mogelijk te benaderen, bouwt de historicus een « *stemma codicum* » (of genealogie van de gegevens), een techniek die ontleend is aan de filologie. In alle gevallen berust de reconstructie op gissingen. Een ander, recenter voorbeeld : in 2000 werd een programma opgezet onder de bescherming van de Nationale Bibliotheek van Berlijn met het oog op de restauratie van de originele partituren van het werk van Johan-Sebastian Bach. De componist maakte namelijk gebruik van een ijzerhoudende inkt die bij oxidatie een zuur vrijgeeft dat zeer agressief is voor het papier en bepaalde partituren geleidelijk onleesbaar

¹ Deze Techno is een bewerking van een artikel over hetzelfde thema : Boydens I., La conservation numérique des données de gestion. *Revue Document Numérique* (Speciaal nummer : « *Archivage et pérennisation* »), Parijs : Editions Hermès Science, (is verschenen in september 2004).

² Chartier R. et Martin H.-J., eds, Histoire de l'édition française. *Tome I. Le livre conquérant*. Du Moyen-âge au milieu du XVIIème siècle, Parijs, Fayard-Cercle de la Librairie, 1989.

maakt³. We dienen evenwel te erkennen dat informatie gedurende eeuwen op « papieren » drager bewaard kon worden terwijl de levensduur van de digitale informatie, als men niet oplet, veel korter kan blijken.

1.2. De bewaring van digitale gegevens

De problematiek is complexer wanneer het om digitale gegevens gaat. De reeksen van bits waaruit ze bestaan, zijn niet « vanzelfsprekend », in tegenstelling tot een tekst op papier of op een muur waartoe de gebruiker « directe » toegang heeft (als hij de taal ervan begrijpt; denk maar aan de hiërogliefen die pas in het begin van de 19e eeuw ontcijferd werden door Champollion). Digitale informatie is kwetsbaar, of zij nu « *born digital* »⁴ is of het resultaat is van een digitalisering⁵. Een voorbeeld : het Amerikaanse departement van Defensie diende een dure « migratie » op te zetten om bestanden te restaureren met betrekking tot de geografische coördinaten van de bommen die gegoooid werden tijdens de Vietnam-oorlog. Deze gegevens, die indertijd opgeslagen werden in een « eigen » database, waren onleesbaar geworden : enkele jaren later was er op de markt geen enkele software meer te vinden waarmee men ze kon verwerken zonder inconsistenties te genereren. De inzet was aanzienlijk want het ging om de ontminning en identificatie van bommen die tijdens of na de oorlog niet ontploft waren en dus nieuwe slachtoffers konden maken. De inconsistenties konden verbeterd worden dankzij de diensten van de « *National Archives* » die het originele eigen formaat bewaard hadden en het periodiek overgezet hadden naar modernere omgevingen, zowel qua software als qua formaat⁶. Digitale informatie berust aldus op een complexe « *chaîne de médiation à la fois matérielle et logicielle* »⁷.

In het punt 2 hierna definiëren wij twee belangrijke componenten van de problematiek : welke informatie wil men bewaren en voor hoelang ? Punt 3 gaat in op de factoren die aan de oorsprong liggen van het verval van de digitale informatie (veroudering van de hardware, van de fysieke dragers, van de softwarecomponenten en de formaten en ook de onderlinge invloeden tussen deze factoren). Daarna evalueren we (punt 4) de vandaag voorgestelde strategieën inzake bewaring (« refreshing », migratie, technologiebehoud, emulatie, inkapseling en gebruik van metadata). Deze traditioneel aangehaalde technieken vullen we aan met een ander punt : in de context van het beheer van « levende »⁸, administratieve toepassingen zullen wij enkele bijkomende pistes aangeven die toegepast kunnen worden van bij de creatie van de gegevens, met het oog op een vereenvoudiging van hun latere bewaring, eens zij gearhiveerd zijn, op het einde van hun « levenscyclus ».

Tot besluit (punt 5) vatten wij de problematiek samen en suggereren wij verschillende methodologische aanbevelingen die afgestemd zijn op de context van de sociale zekerheid en, breder gezien, de federale administratie.

³ Het project, waarvan de duur op drie jaar geschat werd, kreeg een budget van één miljoen euro. De documenten werden nageemaakt aan de hand van een kopieertechniek : « *hiervoor wordt de partituur tussen twee met gelatine ingestreeken bladen geperst die korrel voor korrel een afdruk maken van het manuscript. De « negatieven» worden dan vastgezet op een nieuw blad* ». *Programma om de partituren van Bach te redden van zelfvernietiging*, AFP (bericht), 25/02/2000. Zie ook :

<http://www.rtf.be/matieregreise/emissions/mg23/textes/ailleurs.html> (geconsulteerd op 18 maart 2004).

⁴ Dit wil zeggen, direct onder elektronische vorm aangemaakt, via een tekstverwerker of een systeem voor databasebeheer, bijvoorbeeld.

⁵ Een papieren document dat bijvoorbeeld omgezet is in een digitaal bestand door gebruik te maken van optische karakterherkenning.

⁶ Ruggiero A. (éd.), « Preservation of Digital Memory. Risks and Emergencies. Six Case Studies », *The Future of Digital Memory and Cultural Heritage*, Istituto Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche, Florence, 16 en 17 oktober 2003, p. 35-36.

⁷ Masanes J., « L'information technique nécessaire à la préservation à long terme des documents numériques », *International Preservation News, Newsletter of the IFLA Core Activity on Preservation and Conservation*, nr 29, mei 2003, p. 12.

⁸ Onder "levend" verstaan wij gegevens die nog niet aan het einde van hun levenscyclus gekomen zijn en nog steeds geëxploiteerd worden in het kader van operationeel beheer.

2. Voorwerp en duur van de bewaring

Twee vragen komen dadelijk naar voor. Wat wil men bewaren ? En voor hoelang ?

De archivering van de gegevens, en dus hun bewaring, valt onder het « *record management* ». De Engelse term « *record* » omvat elk type document « *produced or received by a person or organisation in the course of business and retained by that person or organisation* »⁹, ongeacht het formaat, de drager of de min of meer gestructureerde aard van de informatie¹⁰. De Europese norm terzake voegt toe : « *A key feature of a record is that it cannot be changed* ». Verder hebben « *records* » in het algemeen een wettelijke waarde¹¹. In deze context beperkt het voorwerp van de bewaring zich tot documenten die aan het einde van hun « levenscyclus » gekomen zijn en waarvan de inhoud vastligt en niet meer gewijzigd zal worden.

In dit artikel breiden wij het veld uit naar de « levende » beheersgegevens, vatbaar voor wijziging omdat, zoals we zullen zien, de problematiek van de bewaring op lange termijn van digitale informatie ontstaat op het moment T_0 , bij de creatie van de informatie.

Wij stellen voor de bewaringsstrategieën voor te stellen in functie van de status van een gegeven in zijn levenscyclus. Binnen de informatie waarvan de bewaring nuttig geacht wordt, kan men de « korte tijd » van de beheersgegevens, de « tussentijd » van de wettelijke bewaring en de « lange tijd » van de historische bewaring onderscheiden :

- De administratieve « beheersgegevens » kaderen in een « korte tijd » tijdens dewelke « *records* » gevormd worden : bijvoorbeeld, wanneer een officiële brief wordt aangemaakt na de extractie van informatie afkomstig uit een database en na afloop van een validatiefase via « *workflow* ». In dit stadium hebben de documenten een directe operationele impact op de waargenomen werkelijkheid (de burgers onderworpen aan de federale administratie, bijvoorbeeld). Hoelang duurt deze korte tijd ? Hoelang duurt het heden ? De tijd tijdens dewelke « levende gegevens » hun « beheersfunctie » vervullen : in het administratieve domein is deze tijd theoretisch korter dan de verjaringstermijn¹² (bijvoorbeeld : vijf jaar in de sector van de Belgische sociale zekerheid).
- Vervolgens onderscheidt men een « tussentijd », die de wettelijke verjaringstermijn dekt. De « *records* » worden niet meer gewijzigd maar kunnen een impact hebben op de werkelijkheid en het aanmaken van nieuwe gegevens impliceren, in geval van een proces of een heropening van een dossier. Zo heeft de recente zaak *Enron* in de Verenigde Staten het belang aangetoond van bedrijfsinformatie : sinds 2003 reglementeert de wet Sarbanes-Oxley in de USA de bewaring van de documenten (inclusief de e-mails) voor elke op de beurs gequoteerde onderneming¹³.
- Uiteindelijk hebben we dan de « lange tijd » gedurende dewelke de « *records* » bewaard zouden moeten worden « voor eeuwig », niet meer voor wettelijke doeleinden maar als mogelijke bron voor de historici van de toekomst.

⁹ *MoReq Specification, Model Requirement for the Management of Electronic Records*, IDA (Interchange Data between Administrations) Programme of the European Commission, maart 2001

(<http://www.cornwell.co.uk/MoReq%20Specification%20v5-2.4.doc>, geconsulteerd op 18 maart 2004), p. 7-8.

¹⁰ ook « gegeven » genoemd in de rest van het artikel.

¹¹ *Information and documentation – Records management*. ISO-norm 15489-1, 2001-09-05, 1e editie, p. 9.

¹² De verjaring is een wetsbepaling waarmee men zich kan ontdoen van een verplichting na een bepaalde periode. De verjaringstermijn evolueert zelf in de tijd en ging in 1996 van 3 naar 5 jaar in de Belgische sociale zekerheid. K.B. van 20 december 1996 tot wijziging van het koninklijk besluit van 23 november 1969 tot uitvoering van de wet van 27 juni 1969 tot herziening van de besluitwet van 28 december 1944 betreffende de maatschappelijke zekerheid der arbeiders. *Belgisch Staatsblad*, 31 december 1996. De wetgeving terzake is soms vaag en verschilt van land tot land en sector tot sector. Er zijn sites met referenties hieromtrent. Bv., voor de USA, die van de National Archives and Records Administration (<http://www.archives.gov/index.html>) of in het Verenigd Koninkrijk, die van het Public Records Office (<http://www.pro.gov.uk>).

¹³ Berdot V., Dossier « L'archivage d'e-mails ». Les éditeurs font feu de toutes lois, *OI Informatique*, nr 1756, 13 februari 2004, p. 40.

Wij zullen ons op deze tijdsstructuur baseren met het oog op een evaluatie van de verschillende methodes om digitale informatie te bewaren. We zullen namelijk zien dat het gebruik van de ene of de andere bewaringsmethode meer of minder relevant is in functie van de status van een « *record* » in zijn « levenscyclus ».

3. De factoren die aan de oorsprong liggen van het verval van digitale informatie

Talrijke elementen kunnen de digitale informatie veranderen. We zullen hier niet ingaan op natuurrampen (brand, overstromingen, aardbeving,...)¹⁴ waarvoor goed gekende voorzorgen genomen moeten worden, zoals het elders opslaan van de gegevens. We behandelen evenmin de moeilijkheden die verbonden zijn aan de interpretatie van de informatie¹⁵. We bespreken achtereenvolgens deze punten : veroudering van de « hardware » en de fysieke dragers, veroudering van de softwarecomponenten, kwetsbaarheid of heterogeniteit van de formaten.

3.1. De « hardware » en de fysieke dragers

Hardwarecomponenten (chips, geïntegreerde circuits,...) van een computer gaan mettertijd achteruit. Het kan ook zijn dat de hardware van een computertype of een deel ervan op termijn niet meer ondersteund wordt door de firma die het op de markt heeft gebracht.

Daarbij komt nog de veroudering van de fysieke dragers en hun leesapparatuur.

In verband met de dragers citeren we enkele gevallen :

- magneetbanden die geleidelijk hun magnetische lading verliezen (hun geschatte levensduur varieert tussen 10 en 30 jaar);
- CD-Worms die verslijten in functie van de leesfrequentie (experimenten op basis van artificieel ouder gemaakte CD-Worms bevestigen dat zij honderd jaar zouden meegaan);
- microfilms waarvan de levensduur, als zij op gepaste wijze bewaard worden, in een droge en frisse omgeving, kan oplopen tot 500 jaar. Met deze bijzonder duurzame drager is het evenwel moeilijk om gegevens automatisch te verwerken : de conversie van digitale media naar dit analoge medium wordt weinig toegepast wanneer functionaliteiten voor automatische opzoeking vereist zijn om toegang te krijgen tot een groot volume van gegevens.

De fysieke dragers zijn bovendien « *machine dependent* »; zij vereisen leesapparatuur (samen met een software « *driver* »). Deze elementen evolueren evenwel mee met de markt. Men weet nu al dat bepaalde formaten van magneetbanden weldra niet meer gelezen zullen kunnen worden. Momenteel ondersteunen bepaalde « *notebooks* » geen « *floppy disks* » meer van het type 3,5". Hoelang gaan de cd-romlezers gehandhaafd worden als men weet dat DVD's met een opslagcapaciteit die minstens vier keer groter is, beschouwd worden als « het » medium voor « permanente » opslag ? Vandaar deze « boutade » : « *If you are saving your data to CD-Rom, save a CD-Rom drive as well* »¹⁶.

¹⁴ Bij een ernstig ongeval (een vliegtuig dat neerstort op een gebouw) dat bijvoorbeeld financiële centra treft, komen bedrijven die gespecialiseerd zijn in het restaureren van documenten meteen na de brandweer en politie in actie; door water en vuur beschadigde documenten worden onmiddellijk ingevroren om te vermijden dat de dragers nog meer beschadigd zouden worden door de verdamping. Perte d'information. Trouver la parade, *Archimag*, december 2003/januari 2004, nr 170, p. 26.

¹⁵ Boydens I., « Les bases de données sont-elles solubles dans le temps ? », *La Recherche hors série* (« *Ordre et désordre* »), nr 9, november-december 2002, p. 32-34.

¹⁶ Davis M., « Memories are Precious », *Butler Opinion Wire*, 29 januari 2004, p. 3.

3.2. De softwarecomponenten

Een informaticatoepassing is niet bruikbaar zonder de software waarmee zij aangemaakt werd of een gelijkaardige software. De veroudering van de softwarecomponenten ligt in het feit dat er geen systematische achterwaartse compatibiliteit (« *backward compatibility* »)¹⁷ is tussen hun opeenvolgende versies. Met de technologische vooruitgang volgen deze mekaar op. Bepaalde softwareproducenten bieden garanties inzake deze achterwaartse compatibiliteit, tenminste voor een aantal versies, als de overgang van de ene versie naar de andere gebeurt tijdens een migratie, waarbij echter gegevens kunnen verloren gaan. Maar in vele gevallen is deze migratie niet voorzien, wat zware verliezen van informatie inhoudt (de linken tussen de tabellen van een database gaan verloren, bijvoorbeeld). In andere gevallen wordt bepaalde « eigen » software gewoon niet meer opgevolgd of verspreid (omdat de producerende firma failliet is gegaan of overgenomen werd).

In de praktijk is de problematiek nog complexer want er zijn interacties tussen verschillende softwarelagen en in het bijzonder tussen een toepassingsgebonden software en het « *operating system* » (OS) dat zich op de machine bevindt. De evolutie van de OS-versies kan impliceren dat niet meer alle functionaliteiten van de ene of de andere software ondersteund worden. Verder zijn er interacties tussen de evoluties van de OS'en en die van de randapparatuur (een printer, bijvoorbeeld), waardoor er een incompatibiliteit kan ontstaan tussen een OS-versie en bepaalde randapparatuur. Men ziet verder interacties tussen de softwareversies en de types van dragers, zoals wat betreft de lezing van de Lotus-Notes-databases die opgeslagen worden op cd-rom en waarvoor een specifieke procedure gevolgd moet worden¹⁸.

We kunnen daaraan toevoegen dat deze gegevensverliezen des te belangrijker zijn naarmate de structuur van de informatie complexer is : « *The more complex the digital resource, the greater the potential loss is likely to be. For example, interchanging the data held in geographical information system (GIS) databases and groupware databases could involve the loss of thousands of links that have taken years of effort to create and which represent the bulk of the value of the database.* »¹⁹. Idealiter zou men moeten beschikken over internationale standaarden om de homogeniteit van de interacties tussen deze verschillende componenten te garanderen. In de praktijk zijn deze evoluties vaak niet te voorzien en hangen zij af van de actoren op de markt.

3.3. De formaten en de coderingstypes

De formaten en de coderingstypes zijn nauw verbonden met de functionele aspecten die in het vorige punt behandeld werden. Om de duurzaamheid te garanderen van de informatie waarvan zij de structuur vormen, moeten de formaten beschreven worden door publieke normen die worden gedocumenteerd en idealiter gecertificeerd door erkende instellingen (Mercuri, 2003). In de praktijk stellen zich twee soorten moeilijkheden :

- enerzijds gebeurt het dat instellingen afwijken van een standaard, waardoor een *de facto* heterogeniteit gegeneerd wordt;

¹⁷ Deze compatibiliteit betekent dat het mogelijk is om met een software van een versie v+1 een toepassing te verwerken die aangemaakt is met de versie v van diezelfde software.

¹⁸ « Also, be aware that Notes databases on a CD can only be viewed by the same major version of Notes that the Notes database was indexed with. The view index and full text index are improved with each major version, i.e., R5 can't read R4 DB on CD, etc. To work around this with old CDs you may still want to view, you can copy the .nsf file to your local system, follow the procedure above, then copy all the files onto a new CDR because CDRW drives are so inexpensive now ». <http://www.keysolutions.com/NotesFAQ/howcd.html>, geconsulteerd op 13 juli 2004.

¹⁹ Feeney M. (éd.), *Digital Culture : Maximising the Nation's Investment : a Synthesis of JISCO/NPO Studies on the Preservation of Electronic Materials*, Londres, National Preservation Office, 1999, p. 45.

- anderzijds verspreiden bepaalde informaticabedrijven van in het begin eigen formaten : de gebruiker is dan afhankelijk van de evolutie van hun versies en van de financiële gezondheid van het bedrijf.

Deze fenomenen, die vrij pregnant zijn op het niveau van de multimediamarkt (*MoReq Specification*, 2001), genereren een « vluchtigheid » van de formaten en verhogen het risico op veroudering van de informatie. Als antwoord op deze tendens biedt het gebruik van het XML-formaat van het W3C zeer interessante perspectieven : de norm zorgt voor eenmaking en biedt het voordeel dat structuur en inhoud gescheiden worden. Verder heeft de « de facto » en « niet-officiële » PDF-standaard zich een weg gebaand doorheen de wereld dankzij zijn openheid en de bijbehorende gratis leessoftware²⁰.

De soorten coderingen zijn al even fundamenteel. Op dit vlak zorgde de opkomst van Unicode en de norm ISO/CEI 10646 inzake de codering van tekens voor een aanzienlijke uitbreiding van het spectrum van de mogelijke talen en zo kon de verwerking en de homogene uitwisseling van informatie over de wereld bevorderd worden²¹.

4. De bewaringsstrategieën

We gaan achtereenvolgens in op de volgende bewaringsstrategieën : « refreshing », migratie, « technology preservation », « emulatie en inkapseling », gebruik van metadata en, tot slot, enkele principes inzake ontwerp en beheer. Deze strategieën vullen mekaar aan : afzonderlijk bekeken zou geen van hen de onder punt 3 besproken problematiek volledig kunnen oplossen. In de besluiten (punt 5) zullen we in een overzichtstabel het verband weergeven tussen de verschillende factoren die aan de oorsprong liggen van het verval van de digitale informatie en de strategie(ën) om dit probleem te verhelpen.

4.1. « Refreshing »

« *Refreshing* » bestaat erin de informatie te kopiëren van een fysieke drager naar een andere, meer recente (bijvoorbeeld, van een magneetband naar een andere of van een optische drager naar een andere). Met deze aanpak kan de kwestie van de veroudering van fysieke dragers opgelost worden.

4.2. Migratie

Migratie bestaat erin de gegevens van een verouderende « *hardware/software* »-configuratie via een programma te converteren naar een andere, meer recente configuratie. Omwille van de veelvuldige interacties tussen softwarelagen kan de methode evenwel leiden tot informatieverlies dat soms moeilijk geïdentificeerd kan worden. Een migratie vereist dan ook integriteitstesten *a posteriori*, samen met intellectuele input, om de consistentie van de gemigreerde gegevens na te gaan ten opzichte van hun voorgaande status. Migratie wordt op grote schaal toegepast en vult de « refreshing » aan. Het is de standaardoptie die gehanteerd wordt in talrijke bedrijven, administraties of instellingen die belast zijn met het beheer van « levende » of gearhiveerde gegevens.

Deze procedure impliceert dat de wijzigingen van softwareversies opgevolgd moeten worden van zodra de verwerkte gegevens aangemaakt zijn. Deze opvolging omvat verder de toegelaten interacties tussen software, *operating systems* en *drivers* van randapparatuur. Wat

²⁰ Haut H., *Memo PDF*, Interne studie, SmalS-MvM, 14 september 1999.

²¹ André J. en Hudrisier H., eds, « Unicode, écriture du monde ? », *Revue Document Numérique*, vol. 6, nr 3-4/, 2002.

de « marktleiders » betreft, worden in het algemeen migratieprogramma's²² geleverd. Men mag zeker geen belangrijke versie « overslaan » aangezien de achterwaartse compatibiliteit slechts gegarandeerd wordt voor een bepaald aantal versies. In alle gevallen kan de bewerking delicaat zijn zoals erkend wordt door de softwareproducenten zelf: « ... *but you also know that upgrading your databases and applications from currently installed Oracle products can be a complex and nerve-wracking job* »²³.

In een zeer constructieve optiek moedigen sommigen de oprichting aan van diensten die de formaten – die integraal deel uitmaken van de elementen onderworpen aan de migratie - documenteren doorheen hun versies²⁴. Aangezien de bewaring op digitale drager problemen stelt, zou men kunnen overwegen om deze documentatie parallel te bewaren onder een papieren formaat. Het is overigens nuttig om te verwijzen naar de certificatie-instellingen terzake (figuur 1).

Name	Abbr.	URL	Coverage	Composition
American National Standards Institute	ANSI	www.ansi.org	Validates and promotes U.S. standards	Private, non-profit
British Standards Institute	BSi	www.bsi-global.com	British Standard 7799 for security policies and audits, also InfoSec certification	Business consortium
Institute of Electrical and Electronic Engineers Standards Association	IEEE-SA	standards.ieee.org	Over 140 security specific standards	IEEE membership
International Information Systems Security Certification Consortium	(ISC) ²	www.isc2.org	Administers CISSP program for security professionals	Private, non-profit
International Organization for Standardization	ISO	www.iso.ch	International codes for standards management	National standards organizations
Internet Architecture Board	IAB	www.iab.org	Protocol standards for the Internet	IETF membership
National Institute of Standards and Technology	NIST	csrc.nist.gov	Computer security resources and standards for unclassified government data	U.S. government
National Security Agency	NSA	www.nsa.gov	Similar to NIST but for classified data	U.S. government
Organization for the Advancement of Structured Information Standards	OASIS	www.oasis-open.org	E-business standards	OASIS membership
Underwriter's Laboratories	UL	www.ul.com	Product safety testing and certification	Private, non-profit
World Wide Web Consortium	W3C	www.w3.org	Web-related interoperable technologies	W3C membership

Table. Computer security-related standards and certification organizations.

Figuur 1. Enkele standaardisatie-instellingen (bron : Mercuri T. R., « Standards Insecurity », Communications of the ACM, december 2003, vol. 46, nr 12, p. 11-19.).

²² Een oorzaak voor migratie kan te maken hebben met de verhoging van de toegelaten grootte van de gegevensbestanden. Men is in bepaalde gevallen overgegaan van een maximumaantal records per bestand van 16 miljoen naar 2 miljard, wat een verhoging van de grootte van de index-entries (die de primaire sleutels bevatten) noodzakelijk gemaakt heeft : migratie is erop gericht de waarden van de oude index-entries om te vormen om ze compatibel te maken met de nieuwe. In andere gevallen wordt de migratie-utility direct geïntegreerd in de nieuwe versie en zijn de functionaliteiten hiervan niet expliciet. Bij gebrek aan een utility en achterwaartse compatibiliteit is het aan de gebruikers om de inconsistenties tussen versies te ontdekken en deze in de mate van het mogelijke te verhelpen.

²³ Burke B., *Inside Oracle database 10g. The Great Migration Week experiment*, december 2003 (http://otn.oracle.com/pub/news/burke_10g_testing.html), geconsulteerd op 27/02/2004.

²⁴ Masanes J., *Op. cit.*, p. 16. Hoewel nog niet volledig, gaat de Internet Assigned Numbers Authority (IANA) in deze richting (<http://www.iana.org/assignments/media-types/>, geconsulteerd op 18 maart 2004)

4.3. « Technology preservation » of computer musea

Sommigen stellen voor alle vormen van *hardware* en *software* die vandaag gebruikt worden om de gegevens te verwerken, op één plaats te bewaren²⁵. De aanpak is niet realistisch : door de bewaring op één enkele plaats verdwijnt de notie van netwerk, die cruciaal is in de informatica. Zo zouden gebruikers die hun oude informatie wensen te lezen, met hun gegevens naar deze « technologische Ark van Noach » moeten trekken. De kosten qua ruimte, onderhoud van oude uitrustingen en « *knowhow* » van de technici zouden bovendien onbetaalbaar zijn.

4.4. Emulatie en inkapseling

Emulatie²⁶ berust op de logische inkapseling van de gegevens in een « documentaire » laag die de originele omgeving beschrijft (« *hardware/software* » en contextueel) waarin ze aangemaakt werden. Het doel is om dan, op deze basis, de oorspronkelijke omgeving opnieuw aan te maken en ze te emuleren (het gedrag ervan te imiteren) op een nieuwe platformgeneratie. Een experiment in de Koninklijke Bibliotheek van Nederland lijkt succesvol te zijn maar er werden nog onvoldoende testen gedaan : « *de emulatie zorgt voor belangrijke technische problemen (een informaticaplatform « simuleren » is zeker geen triviaal probleem)* »²⁷. De methode wordt namelijk beschouwd als risicovol en krijgt kritiek omwille van het gebrek aan algemeenheid : « *Anyone relying solely on this strategy could be taking a significant risk. They would depending on the technical ability to the software engineers to emulate a specific environment and sustain it, and on the commercial viability of anyone providing such a service.* »²⁸. « *Large-scale, long term emulation is not well-understood. It also has risks of loss of functionality and other characteristics* »²⁹.

4.5. Gebruik van metadata

Verschillende standaarden inzake metadata³⁰ werden voorgesteld om de voorgaande aanpak te ondersteunen (« emulatie en inkapseling ») en, op bredere schaal, bij te dragen tot het globale begrip van de software- en hardwarematige omgeving van digitale informatie. In hun hoedanigheid van instrument voor informatie-interpretatie vereenvoudigen metadata de bewaring van deze informatie : zij laten toe de bewerkingen inzake « refreshing » te documenteren, of migraties uit te voeren op basis van de kenmerken (softwarecode, formaat, enz.) van de oorspronkelijke gegevens die gemigreerd moeten worden. In die zin zijn metadata nuttig ongeacht de voorgestelde bewaringsmethode.

Het XML-formaat biedt het voordeel dat er een scheiding is tussen de waarden en hun logische structuur, ongeacht het software- of hardwareplatform, en wordt hiervoor dan ook meer en meer aanbevolen. Bij de normen voor metadata vindt men bijvoorbeeld het OAIS³¹

²⁵ Kleinberg K. en Logan D., « Digital Preservation in Healthcare : Long-Term Accessibility ». *Gartner Research Note, Strategic Planning*, SPA-15-0907, 7 januari 2002.

²⁶ Rothenberg J., *An experiment in Using Emulation to Preserve Digital Publications*, Den Haag, Koninklijke Bibliotheek, 2000.

²⁷ Masanes J., *Op. cit.*, p. 13.

²⁸ Feeney M. (éd.), *Op. cit.*, p. 42.

²⁹ *MoReq Specification, Model Requirement for the Management of Electronic Records*, IDA (Interchange Data between Administrations) Programme of the European Commission, maart 2001 (<http://www.cornwell.co.uk/MoReq%20Specification%20v5-2.4.doc>, geconsulteerd op 18 maart 2004), p. 85.

³⁰ Het woord « metadata » komt van het Griekse « meta » dat betekent « op », « boven ». Het kan verwijzen naar gegevens die een ander gegeven van lager niveau beschrijven met de bedoeling de interpretatie ervan te vereenvoudigen. Bijvoorbeeld, in verband met een gegeven « datum », kan een metagegeven voorkomen onder de vorm van een zin in natuurlijke taal die specificeert dat het gaat om de « facturatedatum in een boekhoudstelsel ».

³¹ <http://ssdoo.gsfc.nasa.gov/nost/isoas/>

(Open Archive Information System), ontwikkeld op initiatief van het CCSDS (*Comité Consultatif pour les Systèmes de Données spatiales*), ISO-standaard (ISO 14721 :2002). De norm stelt enerzijds weergave-informatie voor waarmee het digitale document en de aanmaakwijze hiervan beschreven kunnen worden en anderzijds informatie die beschrijft op welke manieren het gedigitaliseerde object bewaard wordt³².

Hoewel metadata noodzakelijk zijn³³, vormen zij evenwel niet « de » mirakeloplossing en dit om twee redenen.

Enerzijds zijn de metadata gegevens. Als digitale gegevens niet duurzaam zijn, waarom zouden metadata dat dan wel zijn ? Terwijl de normen voor digitale bewaring het begrip te bewaren digitaal gegeven (of « record ») definiëren, wordt het onderscheid tussen gegevens en metadata soms omzeild : « *the distinction between data and its metadata can be unclear... these details of metadata usage are beyond the scope of the MoReq specification* »³⁴. Sommige studies beschouwen impliciet dat de metadata duurzamer van aard zijn dan de gegevens waarvan zij geacht worden de bewaring te vereenvoudigen. De metadata ontsnappen echter niet aan de overvloed van standaarden : naast de norm OAIS vindt men bijvoorbeeld de norm METS (*Metadata Encoding and Transmission Standard*) : passen deze standaarden in mekaar zoals « Russische poppetjes »³⁵ ? In werkelijkheid stelt de vraag van de duurzaamheid van fysieke dragers, software en formaten zich op dezelfde manier wat betreft de gegevens en de metadata.

Anderzijds botst het beheer van de metadata op drie mogelijke hinderpalen. De eerste hangt samen met het feit dat deze systemen oneindig uitbreidbaar zijn. De meta-informatiesystemen zijn namelijk bestemd om de ondoorzichtigheid van de formele coderingen te verduidelijken en de onzekerheid hiervan te verminderen. Daartoe ligt de vereiste oplossing vaak in het gebruik van een rijkere en dus explicietere codering : de natuurlijke taal³⁶. De natuurlijke taal is echter haar eigen meta-taal. Elke omschrijving die onder deze vorm wordt uitgebracht, kan dan ook het voorwerp uitmaken van een oneindig aantal commentaren van hogere orde. Dit wordt vertaald door de zwaarte en de kostprijs van hun praktisch beheer, wanneer dit steunt op een manuele bijwerking. De tweede hinderpaal is het feit dat de metadata zelf verkeerd en onzeker kunnen zijn : hun validatie kan niet systematisch het voorwerp uitmaken van strikte integriteitstesten. De derde hinderpaal hangt samen met het tijdsverschil tussen de bijwerking van een gegeven en van het bijbehorende metagegeven, waarbij dit laatste, vooral als het voorkomt onder tekstuele vorm, meestal pas aangemaakt wordt op het einde van een min of meer lange analysefase.

In dit opzicht vormen de aanbevelingen van het « Semantisch web », ontwikkeld onder de bescherming van het W3C, een opportuniteit voor de toekomst: zij streven ernaar de informatie te beschrijven dankzij normen die geëxploiteerd kunnen worden door een machine en die begrijpelijk zijn voor mensen³⁷. Over verschillende toepassingsgebieden heen zouden deze normen, waarvan de invoering gebaseerd is op een consensus tussen

³² Masanes J., *Op. cit.* p. 14-18.

³³ Talrijke « levende » toepassingen omvatten het gelijktijdig beheer van omvangrijke databases en van de bijbehorende metadata, zoals, in het domein van de sociale zekerheid, de glossaria die de technische documentatie bevatten van de DmfA-kwartaalaangiften en ASR's. BOYDENS I., *E-gouvernement en Belgique : un retour riche d'expériences*. Techno Speciale Uitgave, nr 26, november 2003.

³⁴ *MoReq Specification, Model Requirement for the Management of Electronic Records*, IDA (Interchange Data between Administrations) Programme of the European Commission, maart 2001 (<http://www.cornwell.co.uk/MoReq%20Specification%20v5-2.4.doc>, geconsulteerd op 18 maart 2004), p. 7.

³⁵ Walbel G., « Like Russian Dolls : Nesting Standards for Digital Preservation », *RLG DigiNews*, 15 juni 2003, vol. 7, nr 3 (Web-based Newsletter : <http://www.rlg.org/preserv/diginews/diginews7-3.html>, geconsulteerd op 18 maart 2004).

³⁶ Gesproken taal, die evolutief is en gecodificeerd wordt door het gebruik alleen.

³⁷ Boydens I., Du « Web sémantique » au « Web pragmatique », SmalS-MvM Research Note, maart 2004, nr 5, p. 19

gebruikersgemeenschappen, de creatie kunnen vereenvoudigen van een semantisch web dat gegevens en metadata verbindt³⁸.

Hoewel de metadata onmisbaar zijn, dient men er spaarzaam mee om te springen en de voorkeur te geven aan halfautomatisch aangemaakte metadata³⁹. Bovendien moet men parallel ingrijpen in de « beginfase », wanneer men « greep » heeft op de bron, bij de samenstelling en bij het beheer van de digitale gegevens, zoals wij dit bespreken in het volgende punt.

4.6. Enkele principes inzake ontwerp en beheer

In zekere zin begint de lange termijn op het moment t_0 van zodra de gegevens aangemaakt worden. Enkele praktische principes inzake ontwerp en beheer die van bij de start gehanteerd worden, maken het mogelijk de bewaring van de gegevens te bevorderen :

- in verband met de formaten, keuze van open en publiek toegankelijke specificaties;
- in verband met de ontwerpmodaliteiten van een toepassing, aanwending van het hergebruikprincipe, bijvoorbeeld via de introductie van het concept « WOPM », « *Write Once Publish Many* »⁴⁰, dat erin bestaat een unieke bron aan te maken (in XML) en deze beschikbaar te stellen onder diverse formaten (ASCII, PDF,...). Dit maakt het mogelijk de gegevens homogeen te exploiteren voor diverse doeleinden (zodat de consistentie tussen de transactionele gegevens en de bijbehorende metadata gehandhaafd wordt, bijvoorbeeld wat de coderingen betreft);
- op het vlak van het beheer wordt het volgende aangeraden :
 - de complexiteitslagen vermijden die de migratiefasen bemoeilijken (vercijfering, compressie), wanneer zij niet noodzakelijk zijn;
 - strategieën ontplooiën om de kwaliteit van de gegevens te verbeteren, bijvoorbeeld :
 - continu de performantie van de processen onderzoeken;
 - de afstemming van de gegevens op de gebruikswijzen opvolgen (wat de detectie inhoudt van « ongebruikte », redundante,... gegevens);
 - gekruiste vergelijkingen tussen « *back-ups* » of tussen concurrerende bestanden die gelijkaardige gegevens bevatten (adreslijsten, bijvoorbeeld) afkomstig van verschillende bronnen.

In verband met het derde punt moet, in het geval van inconsistentie, een onderzoek gevoerd worden naar de oorsprong hiervan (verandering van een drager, bijvoorbeeld) en moet een nieuwe « integere » versie aangemaakt worden. In andere gevallen getuigen de inconsistenties van het feit dat het informatiesysteem en de bevatte werkelijkheid conceptueel niet op mekaar zijn afgestemd (door een evolutie van de werkelijkheid) en moet de structuur van de toepassing herzien worden.

³⁸ Day M., « Integrating Metadata Schema Registries with Digital Preservation Systems to Support Interoperability : a Proposal » 2003 Dublin Core Conference, Seattle, 28 september – 2 oktober 2003 (<http://www.siderean.com/dc2003/Paper38-abstract.pdf>), geconsulteerd op 29 februari 2004.

³⁹ Bijvoorbeeld, op basis van vooraf gecontroleerde lijsten van sleutelwoorden die de betekenis van de velden van een database specificeren.

⁴⁰ Boydens I., « Déploiement coopératif d'un dictionnaire électronique de données administratives », *Revue Document Numérique* (« *Création et gestion coopératives de documents numériques d'information et de communication* »), vol. 5, nr 3-4/2001, p. 27-43.

Van de moeilijkheden die aan de bron liggen van de niet-bewaring van digitale informatie, zijn de vragen rond de evolutie van software en formaten veruit de meest onrustwekkende.

Figuur 2. Overzicht van de interacties tussen de factoren die aan de oorsprong liggen van het verval van digitale informatie.

[illegible]

Figuur 3. De technieken voor de bewaring van digitale informatie.

We hebben bij de aanvang van dit artikel drie « momenten » gedefinieerd in de levenscyclus van de gegevens : de « korte tijd » van de beheersgegevens, de « tussentijd » van de wettelijke bewaring en de « lange tijd » van de historische bewaring. Figuur 4 hieronder vat samen welke strategieën aanbevolen worden in functie van de status van een gegeven in zijn levenscyclus.

⁴¹ Picart K., De la conservation à long terme de l'information numérique, mémoire de fin d'études, Université Libre de Bruxelles, 2000.

	korte tijd (2 jaar)	tussentijd (5 tot 10 jaar)	lange tijd (10 jaar en meer)
"Refreshing"	X	X	X
Migratie	X	X	X
Emulatie en inkapseling			indien rijp
"Technology preservation"			
Metagegevens	X	X	X
Ontwerpprincipes	X		

Figuur 4. De mogelijke strategieën in functie van de levenscyclus van de gegevens.

Een combinatie van « refreshing » en migratie is noodzakelijk doorheen de volledige levensduur van de gegevens.

Als het gaat om het beheer van « levende » gegevens die gehanteerd worden « in de korte tijd », dragen aanvullende strategieën bij tot de bewaring van de informatie, zoals bijvoorbeeld het gebruik van toepassingen met een zuinige structuur (volgens het WOPM-principe, onder meer) of de aanwending van strategieën met het oog op de controle van de kwaliteit van de gegevens. Deze kwaliteitsaspecten vereisen een gestructureerde organisatie : het is niet enkel hun doel de integriteit van de informatie te vrijwaren maar ook de bewaring hiervan op lange termijn te verzekeren. Het komt er tevens op aan open en gedocumenteerde normen te kiezen voor de formaten (zoals de XML-norm van het W3C) en de codering.

Hoe eenvoudiger de omgeving van een gegeven is, hoe « duurzamer » dit is aangezien er minder complexiteitslagen verwerkt moeten worden voor de bewaring op lange termijn. Het is bijvoorbeeld zo dat microfilms soms gebruikt worden voor de lange termijn ondanks hun « functionele passiviteit ». Wat de grote volumes aan gestructureerde gegevens betreft, blijft een digitaal medium vereist.

Wanneer het om informatie gaat, is het, gezien de draagwijdte van de problematiek, beter om in alle gevallen te spreken van « continuïum » in plaats van « duurzaamheid ». Het lijkt immers een illusie om de gegevens voor altijd vast te leggen in hun huidige vorm zonder enige verandering. Gegevensverlies kan zich bijvoorbeeld voordoen tijdens migratiebewerkingen, die nochtans noodzakelijk zijn. Men ziet dan ook een vorm van « continuïum » : de bewaarde gegevens evolueren op één of andere wijze in de tijd.