Digitale Archivering in Vlaamse Instellingen en Diensten

DAVID

Standaarden voor digitale archiefdocumenten

Filip Boudrez

FACULTEIT RECHTSGELEERDHEID
INTERDISCIPLINAIR CENTRUM VOOR RECHT EN INFORMATICA
TIENSESTRAAT 41
B-3000 LEUVEN

Stadsarchief Stad Antwerpen

Versie 4.1

Wettelijk depot D/2001/9.213/10

Antwerpen-Leuven, oktober 2001

Website DAVID-project: http://www.antwerpen.be/david

E-mailadres: david@stad.antwerpen.be

INHOUDSTAFEL

I. INLEIDING

Eén van de voornaamste bekommernissen bij digitale archivering is het verzekeren van de leesbaarheid van de informatie op lange termijn. Veel digitale archiefdocumenten hebben namelijk de eigenschap dat ze afhankelijk zijn van de omgeving waarbinnen ze werden gecreëerd. De ICT evolueert voortdurend met als gevolg dat hard- en software snel verouderen of in onbruik raken. Men kan op verschillende manieren met dit probleem omgaan[1], maar de basisvoorwaarde bij elke vorm van digitale archivering blijft de leesbaarheid van digitale archiefdocumenten. De computerbestanden met de status van archiefstuk moeten leesbaar zijn op andere computers dan waarop ze werden aangemaakt. Men gaat er immers van uit dat de oorspronkelijke computerconfiguraties niet op lange termijn bewaard kunnen worden. Tot op heden betekent dit meestal dat het archiefdocument als een uitwisselbaar computerbestand wordt opgeslagen en in het digitaal archief wordt opgenomen. Uitwisselbaar houdt in dat andere computers weten waarvoor de bytes staan, binaire computerbestanden correct verwerken en toegang hebben tot de drager met de digitale archiefbestanden. Op die manier wordt er op een technologisch onafhankelijke wijze gearchiveerd. Men bereikt dit door een beroep te doen op standaarden op het vlak van codetabellen, bestandsformaten en dragers.

Het belang van gestandaardiseerde codetabellen ligt voor de hand. Tekstuele informatie-uitwisseling tussen computers is enkel mogelijk wanneer dezelfde codetabel wordt gebruikt. Zo niet krijgt men andere karakters te zien dan de auteur bedoelde. De letterkarakters worden immers als hexadecimale waarden opgeslagen en men moet weten naar welk karakter de hexadecimale waarde verwijst. Bij de archivering van tekstuele bestanden komt het er niet alleen op aan om een gestandaardiseerde codetabel als basis te nemen, maar ook om expliciet mee te archiveren welke codetabel werd gebruikt.

Computerbestanden bevatten naast tekstkarakters ook binaire tekens die op een correcte manier moeten verwerkt worden. De binaire tekens en de verwerking ervan zijn afhankelijk van het gebruikte bestandsformaat. Die verwerking is veelal platformafhankelijk. Voor archiveringsdoeleinden werd tot nu toe voor binaire bestandsformaten het meest de migratiepiste toegepast. Deze strategie houdt in dat de digitale archiefdocumenten naar een nieuw bestandsformaat of een geschikt archiveringsformaat worden omgezet. Hierbij worden de computerbestanden bij voorkeur naar een standaard gemigreerd. Een bestandsformaat kan ten eerste een standaard zijn omdat een standaardisatieprocedure aan de basis ervan ligt. Deze standaarden zijn platform- en vendoronafhankelijk en worden geïmplementeerd in diverse computerprogramma's. Hierdoor zijn deze standaarden ook geschikt voor de uitwisseling van computerinformatie. In de meeste gevallen is dit overigens de voornaamste reden voor standaardisatie. De afhankelijkheid van softwaretoepassingen neemt bijgevolg af en de kans wordt kleiner dat men computerbestanden moet migreren wanneer één bepaalde applicatie niet meer beschikbaar is. Door standaarden te gebruiken, moet het mogelijk zijn om op zijn minst een aantal migraties uit te sparen. Een bestandsformaat kan ten tweede uitgroeien tot een standaard wanneer het een groot marktaandeel

[1] Zie hierover het DAVID-rapport: *Het digitaal archiveringssysteem: beheersinventaris, informatielagen en beslissingsmodel als uitgangspunt*, p. 7-11.

verovert, normerend wordt en *defacto* de plaats inneemt van een standaard. Deze standaarden zijn soms niet platform- en/of vendoronafhankelijk, maar kennen wel een grote toepassing en veel computerapplicaties zijn ermee compatibel.

De keuze voor gestandaardiseerde dragers wordt door andere redenen verantwoord. Duurzame dragers in termijnen van decennia bestaan nog niet. Alle dragers takelen af zodat het regelmatig overzetten naar een nieuw medium een basishandeling bij het beheer van digitale archiefbescheiden zal blijven. Bovendien evolueren opslagmedia en bijhorende hard- en software heel snel, zodat de marktevolutie tot het gebruik van nieuwe opslagmedia dwingt. Standaardisatie op het vlak van dragers is daarentegen wel belangrijk om ervoor te zorgen dat computerbestanden vanop hun drager op andere computers inleesbaar zijn. Dit is afhankelijk van hardwarecomponenten (o.a. leesapparaten), software (o.a. drivers: I/O-opdrachten) en de ordening van de informatie op de dragers (formattering, bestandssysteem).

Standaarden zijn belangrijk binnen de informaticawereld. Computergebruikers passen onbewust een aantal standaarden toe. Wil men zoveel mogelijk digitaal archiveren op een technologisch onafhankelijke wijze, dan is het gebruik van standaarden des te belangrijker. Veel archiefinstellingen hanteren dan ook een lijst van de standaarden die zij ondersteunen en waarnaar de over te dragen digitale archiefbestanden worden omgezet alvorens ze in het digitaal depot worden opgenomen[2].

Het belang van standaarden moet anderzijds ook wat genuanceerd worden. Er zijn een aantal belangrijke kanttekeningen. Standaardisatieprocedures nemen ten eerste heel wat tijd in beslag. Ze kunnen zelfs jaren duren met als gevolg dat ze niet altijd aansluiten bij de recentste ontwikkelingen. Aan de andere kant wordt hierdoor een stukje stabiliteit gewaarborgd. Officiële standaardisatie is ten tweede niet altijd gewenst door de ontwerpende ondernemingen want hierdoor verliezen ze voor een stuk hun greep op hun produkt. Ten derde breiden softwareproducenten vanwege commerciële redenen standaarden soms uit met extra functionaliteiten. Die uitbreidingen zijn maar zelden uitwisselbaar. Ten slotte dient opgemerkt te worden dat er niet voor elke computertoepassing standaarden voor handen zijn, zodat men soms niet anders kan dan in zekere mate hard- en softwareafhankelijk te zijn.

Dit werkdocument biedt een overzicht van de standaarden die bij digitale archivering gebruikt kunnen worden. Achtereenvolgens komen de standaarden voor codetabellen, bestandsformaten en dragers aan bod. Van elke standaard worden kort de voornaamste eigenschappen beschreven. Waar mogelijk wordt meer informatie gegeven over het gebruik van de standaard voor archivistische doeleinden. De opsomming van de standaarden is hiërarchisch ingedeeld. De officiële standaarden komen altijd eerst aan bod, gevolgd door de defacto standaarden. Officiële standaarden zijn vastgelegd door officiële standaardiseringsorganisaties en danken hun officiële status aan de participatie van een (inter-)gouvernementele organisatie. Bekende voorbeelden hiervan zijn *ISO* (International Organisation for Standardisation), *IEC* (International Electrotechnical Commission), *ITU* (International Telecommunications Union). Daarnaast zijn er veel regionale en nationale officiële standaardiseringsorganisaties[3]. De defacto standaarden zijn vastgelegd door niet-officiële standaardiseringsinitiatieven (bijv. W3C) of danken deze status aan hun wijdverspreidheid. In dit laatste geval gaat het doorgaans om standaarden die eigendom zijn van een bepaalde producent. Binnen deze groep wordt nog een onderscheid gemaakt tussen de open en gesloten bestandsformaten.

---

2 Bijvoorbeeld: http://www.archives.govt.nz/statutory_regulatory/er_policy/chapter_6_frame.html; *Keeping electronic records: Appendix A - Transfer Records to the Custody of the Australian Archives: Technical Requirements* (http://www.naa.gov.au/recordkeeping/er/keeping_er/append_a.html).

3 CEN/ISSS. *Survey of standards-related fora and consortia*, Brussel, 1998, p. 11-19.

Het overzicht is niet volledig. De IT is voortdurend in ontwikkeling zodat continue opvolging en aanpassing nodig is. Bovendien is voor dit document voor een stapsgewijze aanpak geöpteerd. De eerste versie omvatte de codetabellen, de bestandsformaten voor tekstuele bestanden en de standaarden voor magnetische en optische dragers. Voor de tweede versie werd het overzicht met bestandsformaten voor tekstuele documenten verder uitgebreid en werden de bestandsformaten voor tweedimensionele grafische rasterafbeeldingen toegevoegd. In het hoofdstuk over media werden de recente DVD Blu-ray standaard en de High-Density ROM opgenomen. In versie 3.0 werden standaarden voor audio-visuele bestanden toegevoegd en een algemene bespreking over compressietechnieken en het digitaliseringsproces van geluidssignalen opgenomen. Standaarden voor vectoriële afbeeldingen en magnetische tapes werden aan de vierde versie toegevoegd. Vanaf deze versie wordt ook apart aandacht besteed aan PostScript.

Wat u in toekomstige versies nog mag verwachten, kan uit de (tussen)titels worden afgeleid. Aanvullingen, opmerkingen en suggesties zijn altijd welkom en kunnen doorgemaild worden naar filip.boudrez@sd.antwerpen.be

II. STATUS

VERSIENR	DATUM	INHOUD / TOEVOEGINGEN
1.0	01/11/2001	Codetabellen, tekstuele bestandsformaten, magnetische en optische dragers
2.0	01/03/2002	Uitbreiding van overzicht tekstuele bestandsformaten (PDF en ODA), toevoeging bestandsformaten voor tweedimensionele grafische documenten (TIFF, JPEG, BMP, GIF, PNG, EPS, FlashPix, ImagePac), toevoeging van blu-ray DVD-specificatie en de High-Density ROM.
3.0	29/03/2002	Uitbreiding met (de)compressie, de audio-visuele standaarden (MPEG-Video, AVI, Quicktime), het digitaliseringsproces van geluidssignalen, de audiostandaarden (MPEG-Audio, PCM, WAV, AIFF, AU).
4.0	28/06/2002	Uitbreiding met: PostScript, CGM, SVG, DXF, DWG, UDF, ISO-1001, ANSI X3.27, IBM Standard Label en SIDF.
4.1	30/08/2002	Aanpassing TIFF, toevoeging GeoTIFF, aanpassing PNG

III. CODETABELLEN

A. OFFICIËLE STANDAARDEN

A.1 ASCII of ISO-646

De American Standard Code for Information Interchange (ASCII) is een codetabel die is vastgelegd door het American National Standards Institute (ANSI) met de initiële bedoeling om informatie uitwisseling tussen computers mogelijk te maken. Oorspronkelijk kreeg de ASCII-tabel de naam ANSI_X3.4-1968 mee. De ASCII-tabel werd als officiële standaard vastgelegd in de ISO-646 norm (1972). De ASCII- of ISO-646 karakterset is 7-bits. Dit betekent dat voor de registratie van 1 letterkarakter 7 bits worden gebruikt. Hierdoor zijn er 2^7 (128) verschillende combinaties mogelijk. In de ASCII-codetabel zijn dus 128 lettertekens vastgelegd, waarvan 94 afdrukbare karakters. De tekens van 0 tot en met 31 en teken nummer 127 worden immers gebruikt voor besturings- of controletekens. De originele ASCII-tabel werd gebruikt voor personal computers en werkstations en bevat de tekens die nodig zijn om de Westerse talen vast te leggen. Er werden diverse nationale varianten van de ASCII-tabel gemaakt[4]. De laatste wijziging dateert van 1991.

A.2 ISO/IEC-8859

Omwille van de beperkte mogelijkheden om met een 7-bitstabel andere talen dan het Engels vast te leggen werd de ASCII-tabel uitgebreid. De eerste 128 tekens werden van de ASCII- of ISO-646 tabel overgenomen. De uitbreiding werd mogelijk door het gebruik van een 8-bits codetabel waardoor 256 verschillende tekens kunnen worden vastgelegd. Deze karakterset werd vastgelegd in de ISO/IEC-8859 standaard (1987-1989), soms ook wel eens ASCII 8-bits of extended ASCII genoemd. De karakters die worden vastgelegd dekken de meeste Westerse talen, alsook een beperkt aantal Arabische, Hebreeuwse, Griekse en Cyrillische karakters. ISO/IEC-8859 bestaat uit verschillende tabellen die elk voor bepaalde talen is bedoeld. Zo is ISO/IEC-8859-Latin 1 de codetabel die de tekens van de West-Europese talen vastlegt en waarin ook de Westerse nationale varianten van de ASCII-tabel zijn in verwerkt[5]. Elke codetabel van de ISO-8859 heeft dezelfde opbouw: de posities 0-127

---

[4] De bijzondere karakters in de nationale varianten staan doorgaans op de (decimale) posities 35-36, 64, 91-96, 123-126. Deze karakters zijn in de eerste plaats diakritische tekens. Op de andere posities staan dan dezelfde tekens als in de US-ASCII-tabel. Dit zijn de karakters a-z, A-Z, 0-9, de spatie en de tekens ! " % & ' ( ) * + - . / : ; < = > ?.

[5] In Latin 1-versie van de ISO 8859 tabel worden de posities 128-159 voor niet-afdrukbare controletekens gebruikt. In de karakterset die Windows gebruikt (WinLatin, Windows code page 1252) worden een aantal van deze posities toch voor afdrukbare tekens gebruikt (bijv. het copyright teken). De codetabellen die DOS-computers gebruiken, worden *code pages* genoemd. Eén van de meest gebruikte is *code page 850*. Deze codetabel bevat dezelfde karakters als ISO 8859-1, maar gebruikt soms andere posities.

bevatten de ASCII-karakters, de posities 128-156 bevatten controletekens en de posities 160-255 worden dan gebruikt voor de karakters van de taal waarop de codetabel zich richt. De karakters van bepaalde talen worden soms in nieuwe tabellen nader gespecificeerd (bijv. de tabellen 1 en 15 voor de Westerse talen)

Tabel 1: De ISO-8859 codetabellen

ISO-8859-1: Latin 1	West-Europese talen: Albanië, Baskisch, Catalaans, Deens, Nederlands, Engels, Fins, Gaelic, Duis, Ijslands, Iers, Italiaans, Noors, Portugees, Spaans en Zweeds	ISO-8859-9: Latin 5	Turks, IJslands
ISO-8859-2: Latin 2	Oost-Europese talen: Albanees, Hongaars, Roemeens, Tsjechisch, Pools, Sloveens, Slovaaks, Kroatisch, Servisch	ISO-8859-10: Latin 6	IJslands, Lets, Litouws, Inuit, Sami
ISO-8859-3: Latin 3	Zuid-Europese talen: Maltees, Esperanto	ISO-8859-11:	Thais
ISO-8859-4: Latin 4	Noord-Europese talen: Lets, Litouws, Groenlands, Laps. Vervangen door ISO-8859-10	ISO-8859-12:	/
ISO-8859-5:	Cyrilisch: Russisch, Bulgaars, Servisch, Macedonisch, Oekraïens	ISO-8859-13: Latin 7	Baltisch: Lets
ISO-8859-6:	Arabisch	ISO-8859-14: Latin 8	Keltisch: Gaelic en Welsh
ISO-8859-7:	Grieks	ISO-8859-15: Latin 9 Latin 0	West-Europese talen met o.a. het euro-teken
ISO-8859-8:	Hebreeuws en Jiddisch		

A.3 ISO-10646 en UNICODE

Op basis van de 8-bits codetabel is het wel mogelijk om de tekens van uiteenlopende talen weer te geven, maar het naast elkaar gebruiken van verschillende codetabellen bemoeilijkte de uitwisseling van computerbestanden. Om hiervoor een oplossing te bieden, startte ISO de ontwikkeling van één grote codetabel in plaats van naast elkaar bestaande codetabellen. Hierdoor kan elk karakter uit elke taal slechts één unieke numerieke waarde krijgen. Alle nationale codetabellen moeten hiervoor in één codetabel worden geïntegreerd. ISO werkte aanvankelijk aan een 16 bits-codetabel waarin 65536 tekens werden vastgelegd. Al vlug bleek een 16 bits-codetabel ontoereikend te zijn, en werd in de mogelijkheid voorzien om een 32 bits-tabel (meer dan 4 miljard tekens) aan te leggen. De 16 bits-codetabel werd in de *ISO-10646: Universal Multiple-Octet Coded Character Set* (UCS) vastgelegd.

De belangrijkste computerbedrijven konden zich niet vinden in de ISO-10646 standaard. Ze verenigden zich in het *Unicode*consortium dat zich tot doel stelde een nieuwe gestandaardiseerde codetabel te ontwerpen. Hun karakterset kreeg de naam Unicode mee en is een defacto standaard. Ondertussen traden de vertegenwoordigers van het Unicodeconsortium toe tot de comités die de ISO-10646 voorbereiden en slaagden ze erin om beide codetabellen op elkaar af te stemmen (bijv. Unicode

2.1 en ISO-10646: 1993; Unicode 3.0 en ISO-10646:2000). Sinds 1991 is er ook samenwerking tussen de werkgroepen van beide initiatieven. Die samenwerking resulteerde in een *Basis Multilingual Plane* (BMP, 16 bits). Hiervoor wordt een twee octet coderingsschema (UCS-2, 16 bits) gebruikt. De 32 bitsversie (UCS-4) maakt gebruik van een vier octet coderingsschema, maar is op heden nog niet beschikbaar.

In tegenstelling tot de ISO standaard(en) is de Unicode karakterset wel vrij en gratis beschikbaar. Unicode is uitgebreider dan ISO-10646 doordat het consortium zich ook bezig houdt met de implementatie van hun karakterset zodat deze probleemloos op verschillende platformen functioneert en gemakkelijker tussen verschillende applicaties kan worden uitgewisseld.

De eerste 256 karakters (0 tem 255) zijn identiek aan de ISO-8859 (Latin one)-codetabel. Bij omzetting van ISO-8859-1 (8 bits) naar Unicode (BMP, 16 bits) verdubbelt de omvang van de bestandsgrootte.

A.4 Andere ISO-codetabellen

ISO-2022-JP: Japanse karakters

ISO-2022-KR: Koreaanse karakters

ISO-2022-CN: Chinese karakters

ISO-6861 (1996): Glagolitische karakters

ISO-9036 (1987): Arabische karakters

ISO-10585 (1996): Armeense karakters

ISO-10585 (1996): Georgische karakters

ISO-11822 (1996): Arabische karakters

ISO-13868 (1995, 2001): Bijkomende karakters voor Europese talen

B. DEFACTO STANDAARDEN

B.1 Unicode

Zie A.3 ISO-10646 en Unicode

B.2 EBCDIC

EBCDIC (Extended Binary Coded Decimal Interchange Code) is de codetabel die door mainframecomputers wordt gebruikt. Mainframes werken immers niet met op ASCII gebaseerde codetabellen. EBCDIC is een 8-bits karakterset die door IBM is vastgelegd. EBCDIC is een

uitbreiding van de 4-bits Binary Coded Decimal codetabel. Net zoals bij ASCII bestaan er diverse (nationale) versies van de EBCDIC-codetabel. Er bestaan eveneens mainframetoepassingen die met een eigen EBCDIC-codetabel werken. Niet alle EBCDIC-karakters komen ook in de ASCII-tabel voor. Er is een wel *International Reference Version* die alle ASCII-karakters bevat, maar de karakters hebben niet dezelfde hexadecimale waarde als in de ASCII-tabel. Bovendien volgen de karakters A tot Z niet onmiddellijk na elkaar. Recent werd het euro-teken aan veel EBCDIC-codetabellen toegevoegd. De Unicode karakterset bevat wel alle Unicode-karakters.

IV. BESTANDSFORMATEN

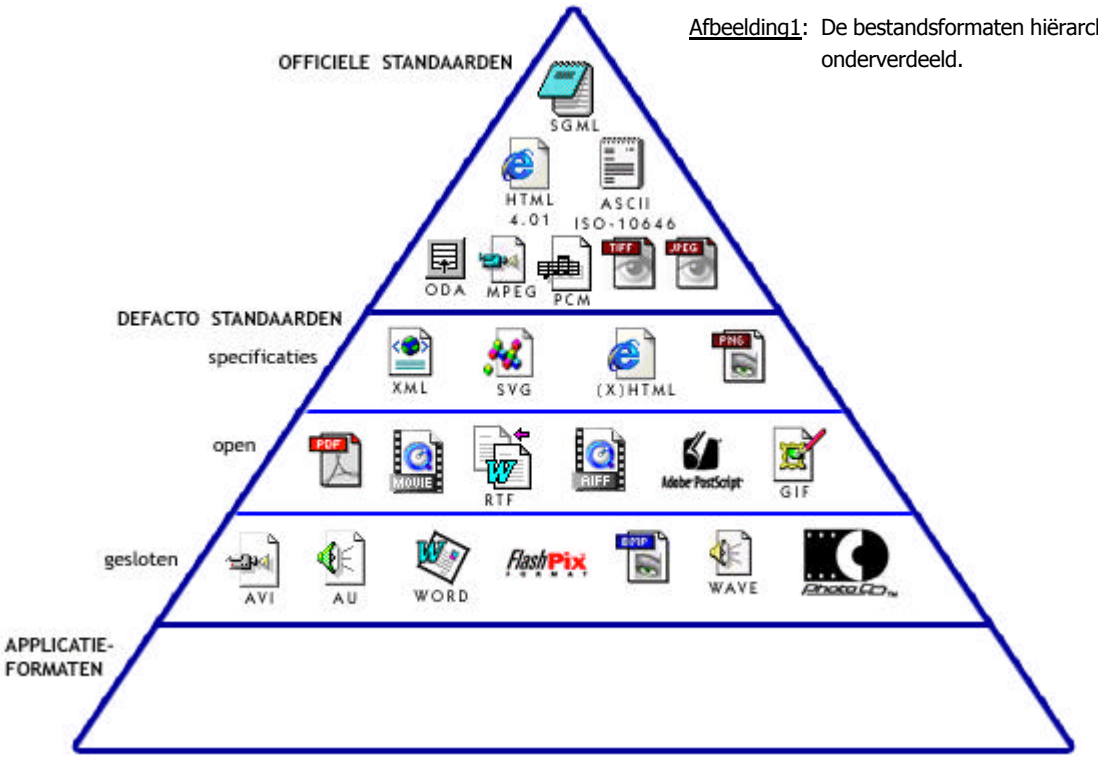
A. ALGEMEEN

De bestandsformaten kunnen hiërarchisch worden ingedeeld. Het criterium voor de indeling die hier wordt gehanteerd, is hun status op het vlak van standaardisatie. Bovenaan in de hiërarchie staan de officiële standaarden. De grootste groep standaarden inzake bestandsformaten zijn de defacto standaarden. Deze groep valt uiteen in drie subgroepen: de specificaties of aanbevelingen vastgelegd door normerende organisaties, de open bestandsformaten afhankelijk van één producent en de gesloten bestandsformaten. De specificaties of aanbevelingen zijn het resultaat van samenwerkingsinitiatieven met normering of standaardisatie als doel. Eén van de bekendste voorbeelden op dit ogenblik is *W3C* (*World Wide Web Consortium*). Net zoals bij officiële standaarden wordt voor het opstellen van de specificatie van deze bestandsformaten een hele procedure gevolgd. Samen met het feit dat meerdere partijen (softwareproducenten, universiteiten, consumenten) bij deze initiatieven zijn betrokken, wordt hierdoor een stukje stabiliteit gewaarborgd. De open en gesloten bestandsformaten hebben als gemeenschappelijk element dat ze eigendom zijn van één producent. Het verschil tussen beide groepen is dat van de open bestandsformaten de specificatie is vrijgegeven en bij de gesloten bestandsformaten niet. In beide gevallen gaat het om bestandsformaten die vanwege hun wijdverspreidheid normerend zijn geworden. Deze bestandsformaten zijn als het ware tot een standaard uitgegroeid. Het onderscheid tussen een al dan niet openbaar gemaakte specificatie is een belangrijk gegeven, want op basis van de gepubliceerde beschrijving kan in de toekomst de nodige programmatuur voor het inlezen van de computerbestanden geschreven worden. Bij gesloten bestandsformaten is dit nauwelijks of niet mogelijk. Helemaal onderaan in de hiërarchie staan ten slotte de bestandsformaten van weinig voorkomende commerciële toepassingen of ad hoc ontwikkelde toepassingen. Deze formaten zijn niet gemakkelijk uitwisselbaar, zijn afhankelijk van één applicatie en kunnen bezwaarlijk een standaard worden genoemd[6].

Bestandsformaten met de status van standaard zijn in de regel geschikte archiveringsformaten voor digitale documenten. Een hiërarchische indeling is van belang om het overzicht te behouden en kan als leidraad dienen bij de keuze van een duurzaam archiveringsformaat. Aan de andere kant is de hiërarchische indeling niet absoluut voor de digitale duurzaamheid op zich. Enige nuancering is op zijn plaats. Inzake leesbaarheid op lange termijn biedt een officiële standaard a priori niet meer garanties dan een defacto standaard. Officiële standaardisatie en marktevoluties lopen niet altijd parallel. Of een bepaalde specificatie in de praktijk echt als een "standaard" kan worden beschouwd, is afhankelijk van de implementatie door softwareproducenten en de toepassing ervan door eindgebruikers. Beide factoren gaan meestal hand in hand. De computerindustrie volgt niet altijd de officiële standaardisatie en consumenten verkiezen doorgaans de meest gebruiksvriendelijke oplossing. De impact van bijvoorbeeld XML versus SGML of van Unicode versus ISO-10646 illustreert dit.

---

[6] Het DLM onderscheidt drie soorten standaarden: defacto, specificaties en de iure standaarden. Een andere veel voorkomende indeling is de opsplitsing in enerzijds de gesloten/producentgebonden formaten ("proprietary") en anderzijds de open of industriële formaten.



Afbeelding1: De bestandsformaten hiërarchisch onderverdeeld.

Het hiernavolgend overzicht is opgesplitst in twee delen: de tekstuele bestanden en de audio-visuele bestanden. Bij elke standaard worden een aantal terugkerende elementen beschreven: de status inzake standaardisatie, de interne structuur van het bestandsformaat, de voornaamste eigenschappen, algemene toepassingen van het bestandsformaat, voorbeelden van archiveringscasussen waarin het formaat wordt gebruikt en de eventuele bescherming door patent- of eigendomsrechten. Dit laatste is van belang voor de eventuele noodzaak om licensies aan te schaffen.

B. TEKSTUELE BESTANDEN

B.1 Officiële standaarden

B.1.1 Platte tekstbestanden

Een plat tekstbestand is een bestand dat enkel uit ASCII- of Unicodekarakters bestaat. Een ASCII-of Unicodebestand bevat geen header of binaire tekens die door applicatiesoftware worden gebruikt. Op de eerste byte van een plat tekstbestand staat een letterteken. In het Engels worden deze bestanden *flat files* genoemd. In de omgangstaal wordt plat tekst-, ASCII- of Unicodebestand gebruikt om het tegenovergestelde van binair bestand aan te duiden.

ASCII-bestanden kunnen door de meeste computerapplicaties voor de verwerking van tekstuele informatie worden ingelezen (teksteditors of -verwerkers, spreadsheetprogramma's, databanksoftware, webbrowsers, enz.). Bij het opslagen van teksten als platte tekstbestanden is het wel belangrijk om vast te leggen welke codetabel en eventueel welke versie werd gebruikt. Bij de toepassing van Unicode is dit overbodig.

In een plat tekstbestand wordt in de eerste plaats de inhoud van een tekstueel computerbestand bewaard. In een plat tekstbestand is het doorgaans ook mogelijk om de structuur van een document te bewaren. Daartoe worden ASCII-karakters als veldscheidingsteken gebruikt. In de meeste gevallen wordt een tab-teken of een punt-komma als delimeter gebruikt. Aangezien deze karakters soms ook binnen gegevensvelden voorkomen, kan dit moeilijkheden opleveren bij latere omzetting of raadpleging. Alle gegevensvelden van één record worden doorgaans op dezelfde lijn geplaatst. De lengte van één lijn is onbeperkt. Het enter-teken wordt als scheidingsteken tussen records gebruikt.

ASCII- of Unicodebestanden bevatten geen afbeeldingen of geluid, maar enkel tekstuele tekens zonder enige opmaakgegevens (onderlijning, lettertypes, puntgrootte, vet, cursief, enz.). Bij de omzetting van een binair tekstbestand (bijv. een Worddocument) naar een plat tekstbestand worden enkel de tekstkarakters bewaard. Alle binaire tekens gaan verloren.

Een plat tekstbestand is software-onafhankelijk. In die zin beantwoordt het volledig aan de archivistische noden. Een belangrijk nadeel is echter dat structuur en opmaak van een tekstueel computerbestand meestal tot de essentiële - en bijgevolg te archiveren - elementen behoort. Wanneer dit het geval is, voldoet een plat tekstbestand niet. SGML en vooral XML kunnen dan een alternatief zijn. Beide formaten zijn uitermate geschikt voor het vastleggen van de structuur. Voor het archiveren van de opmaak kan een stylesheet worden gemaakt.

B.1.2 Standard Generalized Markup Language (SGML)

Standard Generalized Markup Language (SGML) is een markuptaal die de inhoud, de structuur en de semantiek van documenten vastlegt. SGML werd door ISO in 1986 als officiële standaard vastgelegd (ISO-8879). SGML is een metataal die voor de beschrijving van andere markuptalen kan worden gebruikt (bijv. HTML).

In essentie zijn SGML-bestanden platte tekstbestanden waarbij tags als delimeters tussen de gegevensvelden functioneren. De tags leggen de structuur en onderlinge relatie van de elementen vast waaruit het document is opgebouwd. Elk gegevensveld heeft een begin- en eindtag. De tags staan tussen < en >. De eindtag is dezelfde als de begintag met een slash eraan toegevoegd. Het is de bedoeling dat er semantische tags worden gebruikt. De gebruiker kan de tags zelf definiëren. In tegenstelling tot gewone platte tekstbestanden zijn de tags uniek en zeggen ze iets over de inhoud van de informatie in het gegevensveld. Hierdoor worden omzettings- en interpretatiefouten vermeden. Inhoud en semantiek worden samen in één bestand bewaard. Een voorbeeld hiervan is: <publicatiedatum>1 november 2001</publicatiedatum>. Aan de elementen kunnen attributen worden toegekend.

Een SGML-bestand heeft altijd een DTD (Document Type Definition) nodig. In de DTD legt de gebruiker de structuur van het document vast. Hij kan de tags onbeperkt nesten en zo een hele boomstructuur uitwerken. De DTD kan in de header van het SGML-bestand (intern) of in een

afzonderlijk bestand (extern) worden vastgelegd. Bij het openen van het SGML-bestand wordt eerst gecontroleerd of de structuur van het document met de DTD overeenstemt. Dit proces wordt *parsing* genoemd en kan als controlemiddel dienen.

Er zijn geen beperkingen op de lengte van gegevensvelden. Eén gegevensveld kan zowel een cel als een hoofdstuk van een boek zijn. Men kan dus zowel tabellen, databanken of boeken als SGML-bestanden bewaren. In de SGML-bestanden zelf staan geen afbeeldingen, maar de SGML-bestanden kunnen wel verwijzingen naar die afbeeldingen bevatten. Om multimedia- en hyperlinkfunctionaliteiten aan SGML-documenten werd in 1992 een nieuwe standaard vastgelegd door ISO: HyTime (ISO/IEC-10744(1992): *Information technology -- Hypermedia/Time-based Structuring Language*[7]).

Bij SGML ligt de klemtoon op het bewaren van gestructureerde informatie. SGML is heel typisch object-geörienteerd. Hiërarchisch gestructureerde databanken kunnen betrekkelijk gemakkelijk als SGML-bestanden worden bewaard. De databanklogica of -intelligentie kan niet in het SGML-bestand zelf worden gearchiveerd.

Een SGML-bestand op zich bevat geen lay-out. Men kan een SGML-bestand wel een lay-out geven door het aan een stylesheet te koppelen. De stylesheettaal voor SGML is DSSSL (Document Style Semantics and Specification Language). DSSSL is in 1996 vastgelegd als ISO-standaard (ISO-10179). Een stylesheet zal echter zelden op een identieke wijze de lay-out van het oorspronkelijke wijze kunnen weergeven.

SGML gebruikt de ISO-646 codetabel.

SGML is hoofdzakelijk in de uitgeverswereld gebruikt. Veel voorbeelden van SGML-toepassingen voor archiveringsdoeleinden zijn er niet. Dit heeft meerdere redenen. Voor de omzetting van tekstverwerkingsbestanden (WordPerfect, Word) naar SGML zijn wel de nodige programma's op de markt, maar de migraties naar SGML blijven in hoge mate arbeidsintensief. SGML wordt ook nauwelijks of niet toegepast binnen actieve applicaties, zodat er op zijn minst altijd één omzettingsmoment naar SGML nodig is. De omzetting naar SGML vraagt vooral veel werk wanneer de bronbestanden niet op een gestructureerde wijze zijn opgebouwd. Door zijn vele mogelijkheden en flexibiliteit is SGML vrij complex wat zijn algemene verspreiding in de weg stond. Sinds 1998 is er een alternatief voor SGML voor handen: XML. SGML is nooit zo populair geweest als XML nu is.

Referentie: http://www.iso.ch; C.F. GOLDFARB, *The SGML handbook*, Oxford, 1990.

B.1.3 HyperText Markup Language (HTML): 4.01

Hypertext Markup Language (HTML) is de markuptaal voor documenten op het WWW wordt gepubliceerd. Bij een zuivere toepassing van HTML leggen de tags de structurele onderdelen van een document en hun functie vast. HTML-tags definiëren titels, paragrafen, opsommmingen, tabellen, enz. Aan de tags worden attributen toegekend. HTML-bestanden kunnen als plat tekstbestand geopend worden in een gewone teksteditor of tekstverwerker.

HTML wordt vastgelegd door het *World Wide Web Consortium* en heeft algemeen de status van defacto standaard. Dit is ondermeer het geval door de wijdverspreide HTML-versies 2.0 (IETF, 1994),

[7] http://www.iso.ch

3.2 (1996), 4.0 (1997), en 4.01 (1999). Deze laatste versie is als officiële standaard vastgelegd door *ISO* (ISO-15445 (2000): *Information Technology. Document description and processing languages. HyperText Markup Language*). Andere HTML-versies zijn dus geen officiële, maar defacto standaarden. De recentste HTML-versies zijn Dynamic HTML en XHTML. XHTML 1.0 is een herformulering van HTML 4.01 in XML en combineert HTML met de voordelen van XML. XHTML komt in grote lijnen neer op het vertalen van HTML in de XML-syntaxregels. XHTML-pagina's zijn met HTML-browsers compatibel wanneer ze HTML-compatibiliteit regels worden toegepast. Er zijn drie varianten op XHTML vastgelegd: strict (cleane markup in combinatie met CSS), transitional (combinatie met CSS maar met toepassing van een aantal lay out tags en attributen), frameset (toepassen van HTML frames binnen de webpagina). Elke variant heeft zijn eigen DTD. XHTML 1.1 is een gemodulariseerde versie van XHTML 1.0, die het mogelijk moet maken dat gemakkelijk XHTML profielen worden gecreëerd.

HTML is een gesloten markuptaal. Dit houdt in dat de verzameling HTML-tags vast ligt en door de gebruiker niet kan worden uitgebreid. Bij elke HTML versieverhoging wordt het assortiment tags aangepast. Minder populaire tags worden weggelaten of vervangen door nieuwe tags. Of de niet meer ondersteunde HTML-tags nog kunnen uitgevoerd worden, is afhankelijk van de gebruikte webbrowser. De huidige commerciële en wijdverspreide webbrowsers zijn hier behoorlijk soepel in. Ze slagen erin om oude HTML-versies te kunnen inlezen. Er kunnen zich wel problemen voordoen bij niet-gestandaardiseerde tags of verkeerde attributen. Er zijn namelijk een aantal HTML-editors op de markt die bepaalde tags gebruiken die enkel functioneren in de webbrowser van dezelfde producent. Het *W3C* stelt een tool ter beschikking die kan gebruikt worden voor het automatisch opruimen van verouderde en/of afgekeurde tags en attributen.

In een HTML-bestand kunnen inhoud en lay-out samen worden bewaard, maar dit druist tegen de HTML-ontwerpregels in. De HTML-tags en hun attributen zijn niet ontworpen om de lay-out te definiëren. De initiële doelstelling van HTML was de uitwisseling van wetenschappelijke teksten, maar zijn populaireit leidde algauw tot oneigenlijke toepassingen van HTML. Met het oog op toevoegen van lay-out werden nieuwe tags ontworpen (<font>,
, enz.), wat dan weer tot compatibiliteitsproblemen leidde. De huidige tendens is er duidelijk op gericht om beide onderdelen van een document gescheiden te bewaren. De afbeeldingen in een HTML-bestand worden sowieso in een afzonderlijk bestand (bijv. GIF, JPEG, TIFF, PNG) opgeslagen. De HTML-bestanden bevatten enkel een verwijzing naar de afbeelding die op een bepaalde plaats moet worden geopend. Door middel van Cascading Style Sheets (CSS) kan stijl en opmaak aan een HTML-document worden toegevoegd. CSS wordt net zoals HTML door het *World Wide Web Consortium* vastgelegd. Men onderscheidt CSS1 (level 1, december 1996) en CSS2 (level 2, mei 1998)[8]. Net zoals de afbeeldingen zijn stylesheets afzonderlijke computerbestanden.

Als er een einde komt aan de compatibiliteit van webbrowsers met een bepaalde HTML-versie zijn er twee opties. Ofwel zorgt men voor een emulator voor de webbrowser die de HTML-versie ondersteunt, ofwel worden de verouderde tags aangepast. Dit laatste komt in feite neer op migratie. Aangezien er dan niet alleen tags maar ook attributen worden aangepast, houdt dit voor een stuk het herschrijven van de HTML-pagina in. Momenteel wordt er al volop geëxperimenteerd met browseremulatie.

Het bewaren van archiefdocumenten in HTML lijkt niet gebruikelijk te zijn. Gearchiveerde websites worden overwegend in HTML gearchiveerd. HTML kan ook gebruikt worden bij de

[8] Voor meer informatie over CSS, zie: http://www.w3.org/Style/CSS/

archivering van e-mails waarvan de lay-out (doorgaans op basis van een stylesheet en HTML-pagina) belangrijk is. In HTML-pagina's kunnen de metadata als headerinformatie worden opgenomen (via de metatags). Op die manier worden archiefobject en metadata onlosmakelijk aan elkaar verbonden en maken ze deel uit van één en hetzelfde computerbestand.

Referentie: http://www.w3.org/MarkUp/

B.1.4 Open Document Architecture (ODA) and Interchange Format

ODA is vastgelegd door ISO en IEC (ISO-8613: *Information Processing - Text and Office Systems, Office Document Architecture (ODA) and Interchange Format*; ISO/IEC ISP-10610-1:1993; ISO/IEC ISP-11181:1993; ISO/IEC ISP-11182-1:1993; ISO/IEC ISP-12064-1:1995; ISO/IEC ISP-15124-1:1998). ODA bevat een geheel van regels die de uitwisseling van documenten tussen verschillende platformen moet mogelijk maken zonder verlies van inhoud of lay-out. Met documenten worden in de eerste plaats brieven, rapporten en nota's bedoeld. De documenten kunnen in hun logische structuur (processable, hiërarchische beschrijving van de onderdelen), lay-out structuur (formatted, hiërarchische beschrijving van de lay-outobjecten) of een combinatie van beide (formatted & processable) worden vastgelegd. ODA-bestanden kan men in drie niveau's aanmaken: gestructureerde tekst, raster- en/of vectorafbeeldingen en grafieken. In het Document Application Profile (DAP) worden de karakteristieken van een document vastgelegd. ODA gaat heel ver in de beschrijving van de lay-out. Niettegenstaande de steun van de Europese Unie kent ODA maar een kleine verspreiding. De redenen hiervoor zijn een gebrek aan ondersteunende softwareprodukten en de concurrentie van SGML[9]. De Nationale Archiefdienst van Canada startte een pilootproject rond ODA, maar kreeg weinig of geen navolging. Xerox' Raster Document Object (RDO) is grotendeels op ODA gebaseerd, maar is producentgebonden en beschermd.

Referentie: http://www.iso.ch

B.2 Defacto standaarden

B.2.1 eXtensible Markup Language (XML)

F. BOUDREZ, *<XML/> en digitaal archiveren*, Antwerpen, 2002. (http://www.antwerpen.be/David → cases).

Referentie: http://www.w3c.org; http://www.oasis-open.org; http://xml.coverpages.org/;

[9] http://www.infoma.jyu.fi/digimedi/Pasi/eds/odaodif.htm . ODA stond vroeger voor Office Document Architecture.

B.2.2 HTML

Zie B.1.3 HyperText Markup Language

B.2.3 PostScript (PS)

PostScript werd door Adobe gecreëerd en vanaf 1985 verspreid. De specificatie van PostScript wordt vrijgegeven. In een PostScriptbestand wordt beschreven hoe de af te drukken pagina er uit ziet. PostScript is een beschrijvingsmodel gebaseerd op Adobes Imaging Model voor tekst, grafieken en afbeeldingen waarbij de verschijningsvorm in termen van abstracte grafische elementen en niet als apparaatpixels worden gedefinieerd.

PostScript is een apparaat- en platformonafhankelijke beschrijvingstaal waarin een samengestelde tekst naar een raster outputtoepassing (scherm, printer, plotter) wordt gecommuniceerd. Het outputproces bestaat doorgaans uit twee stappen: een applicatie genereert een apparaatonafhankelijke beschrijving van de gewenste output in de paginabeschrijvingstaal en het programma dat een specifiek rasterapparaat aanstuurt, interpreteert de beschrijving en zorgt voor de renditie. PostScriptbestanden kunnen echter ook zonder de tussenkomst van een applicatie naar de printer worden gestuurd. De enige voorwaarde is dat de printer met een PostScriptinterpreter is uitgerust. PostScript is niet gebaseerd op bitmapping zodat de *.ps-bestanden resolutie-onafhankelijk zijn. De omzetting in bitmappatronen gebeurt door het outputapparaat.

Aangezien PostScript in wezen ook een programmeertaal is, bevatten de bestanden leesbare code (ASCII-karakters) die in een gewone teksteditor of tekstverwerker kan worden bewerkt. PostScript maakt een onderscheid tussen hoofd- en kleine letters en alles wat volgt na het procentteken is commentaar. Een goed opgebouwd PostScriptbestand bestaat uit twee delen: een proloog (algemene instructies en proceduredefinities) en het script (de eigenlijke beschrijving van de pagina).

Er is een PostScript Level 1, Level 2 en Level 3. Dit zijn de uitbreidingen op de initiële PostScriptversie. De drie groepen heten officieel Languagelevel 1 tot en met 3, maar worden doorgaans met Level 1, 2 en 3 in applicaties aangeduid. Level 3 is een uitbreiding van Level 2 en bevat de twee voorgaande Levels. In een PostScriptinterpreter (bijv. een printer) die een bepaald Level ondersteunt moeten alle functionaliteiten van dat Level en de voorgaande geïmplementeerd zijn. Een PostScriptinterpreter kan ook bepaalde functionaliteiten ondersteunen die niet tot een bepaald Level behoren, maar die een uitbreiding zijn van een bepaalde applicatie.

PostScript werkt met verschillende compressiefilters (LZW, zlib/deflate, DCT, RLE, CCITT), maar kan evengoed geen compressieloos worden toegepast. Een PostScriptbestand kan zowel een statisch of een dynamisch formaat zijn.

PostScript is bruikbaar voor de uitwisseling of opslag van afdrukbare bestanden die zowel tekst als afbeeldingen bevatten. Op PostScript is PDF gebaseerd. In vergelijking met PostScriptbestanden hebben de PDF-bestanden doorgaans een kleinere bestandsomvang.

Er zijn verschillende computertoepassingen die een PostScriptbestand kunnen samenstellen. Eén van de bekendste voorbeelden is wellicht de Acrobat Distiller.

Referentie: ADOBE SYTEMS, *PostScript Language Reference. Third edition*, 1999.
http://partners.adobe.com/asn/ developer/technotes/postscript.html

B.2.4 Portable Document Format (PDF)

Softwareproducent Adobe startte de verspreiding van PDF midden 1993. PDF is eigendom van Adobe, maar de specificatie wordt vrijgegeven waardoor het een open defacto standaard is. Hoewel Acrobat 5.0 en Illustrator 9.0 al op de specificatie 1.4 zijn gebaseerd, is op de Adobe website enkel de specificatie van versie 1.3 beschikbaar. De specificatie 1.4 is achterwaarts compatibel met de drie vorige versies. Hetzelfde geldt voor de jongste Acrobat-programma's. Het specificatieversienummer staat vooraan als PDF-commentaar in een bestand (bijv. "%PDF-1.2"). Aangezien de specificatie bekend is, kunnen er in principe ook vrij PDF-tools worden ontwikkeld.

PDF is gebaseerd op PostScript (*.ps-bestanden)[10]. Net zoals PostScript is PDF onafhankelijk van de hardware, het besturingssysteem en de applicaties waarmee de documenten werden gecreëerd. Op basis hiervan stelt men PDF als een platformonafhankelijk bestandsformaat voor. Hiermee bedoelt men dat eenzelfde PDF-bestand op verschillende besturingssystemen kan worden bekeken. Voor het openen van PDF-bestanden moet men wel over de viewer Acrobat-Reader beschikken. Met uitzondering van een aantal grafische programma's kunnen andere computerapplicaties PDF-bestanden niet op een leesbare wijze openen. Echt platformonafhankelijk is PDF dus niet. De Readerversie van het Acrobatprogramma kan vrij vanaf de Adobewebsite worden gedownload. Er bestaan Reader-versies voor de verschillende computerplatformen.

Alle gegevens die nodig zijn om documenten in hun originele verschijningsvorm af te drukken of te presenteren, worden in het PDF-bestand zelf bewaard. Hierdoor kan het PDF-bestand worden afgedrukt of op het scherm weergegeven worden zoals de auteur het had bedoeld. Dit maakt de bewaring van de originele "look and feel" van documenten mogelijk. Om dit te realiseren wordt in het PDF-bestand niet alleen het PDF-document maar ook bijhorende ondersteunende data bewaard. De *font descriptors* die voor elk gebruikt lettertype aan het PDF-bestand worden toegevoegd, zijn hier een voorbeeld van. De gebruiker kan bij het bewaren als een PDF-bestand bepalen van welke fonts de metrische en andere informatie (bijv. grootte, dikte, stijl, breedte) in het PDF-bestand zelf wordt bewaard. Dit is aangewezen wanneer er niet-courante fonts bij de pagina-opmaak werden gebruikt. Een aantal fonts die door patenten zijn beschermd, kunnen niet worden opgenomen. Als de ontvanger dan niet over het vereiste font beschikt, dan wordt het ontbrekende lettertype nagebootst op basis van de gegevens in het PDF-bestand zelf. PDF-bestanden zijn bijgevolg binaire bestanden.

PDF-bestanden bestaan uit vier secties: header, body, cross-reference table en trailer. De header bevat het versienummer van de specificatie. De body wordt gevormd door de objecten waaruit het PDF-document is opgebouwd. Elk object is genummerd en wordt gesloten met `endobj`. In de cross-

[10] Er zijn een aantal belangrijke verschillen tussen PDF en PostScript: PDF is geen programmeertaal, PostScript-bestanden kunnen geen hyperlinks bevatten, PDF-bestanden bevatten lettertype metrics. Postcriptbestanden bevatten alle informatie van de documentpagina's, informatie over de gekoppelde bestanden (bijv. geïmporteerde illustraties), letterypen en printerinstructies. Postcript is een geschikt bestandsformaat om bijvoorbeeld PageMakerbestanden op een platformonafhankelijke wijze te archiveren.

reference table (<xref>) wordt informatie opgeslagen zodat er onmiddellijke toegang tot objecten uit de PDF-body is. In de trailer staat de verwijzing naar de startbit van de cross-reference table zodat snelle toegang mogelijk is (<startxref>). Bij het openen van het PDF-bestand wordt immers eerst de trailer ingelezen.

Navigatie binnen het digitale document is mogelijk door thumbnails van pagina's, hyperlinks en bladwijzers. De manier waarop deze informatie in PDF-bestanden wordt opgeslagen, wordt het Adobe imaging model genoemd en is eveneens gebaseerd op PostScript. De basis is steeds een blanco bladzijde. De data bepalen op welke plaatsen er "inkt" van om het even welke kleur komt, welke marges er zijn, welke fontspecificaties worden gebruikt, enz.

PDF-bestanden kunnen tekst met opmaak, grafieken en afbeeldingen bevatten. De tekst in een PDF-bestand kan op twee manieren worden opgeslagen. Enerzijds kunnen tekstuele gegevens rechtstreeks als tekstkarakters worden weergegeven. Dit laat het kopiëren van de tekst naar andere applicaties toe. Anderzijds kan de tekst ook als een afbeelding worden bewaard. Acties zoals zoekopdrachten, kopiëren of omzetting van de tekst naar een ander formaat zijn in dit laatste geval niet mogelijk. De tekst in een PDF-bestand wordt niet als ASCII- of Unicodekarakters opgeslagen[11]. De tekst van een PDF-bestand kan dus niet in teksteditors of in tekstverwerkingsprogramma's worden bekeken. PDF is initieel ontworpen voor 8 bits karaktersets.

PDF-bestanden kunnen twee soorten afbeeldingen bevatten. De pixel georiënteerde afbeeldingen hebben een wiskundige representatie en kunnen relatief gemakkelijk worden gemigreerd. De data van gelinieerde afbeeldingen bepalen hoe elke lijn van de afbeelding er uit ziet. Aangezien deze afbeeldingen niet zo gestandaardiseerd zijn als de pixel georiënteerde kunnen ze niet zo gemakkelijk worden omgezet. Gelinieerd opgebouwde afbeeldingen moeten hiervoor eerst naar pixel georiënteerde worden omgezet. PDF bewaart afbeeldingen op een resolutie onafhankelijke wijze.

Men creëert PDF-bestanden met een Acrobatprogramma van Adobe. PDF-bestanden worden rechtstreeks uit applicaties of uit PostScriptbestanden gemaakt. Een PDF-bestanden kan namelijk op twee manieren worden aangemaakt. De PDF-Writer handelt net zoals een printerdriver. Normaal gezien vertaalt een printerdriver afbeeldingen en tekst naar commando's die printers begrijpen. PDF-Writer stuurt de commando's evenwel niet naar een printer maar converteert de commando's naar PDF operatoren die in een PDF-bestand worden opgenomen. De PDF-Distiller zet PostScriptpagina's in PDF-bestanden om.

De Writer en Distiller comprimeren het bestand. De compressieverhoudingen variëren van 10:1 voor kleurafbeeldingen tot 2:1 voor combinaties van tekst en beeld. Bij het wegschrijven als PDF-bestand kan de gebruiker het compressie-algoritme (bijv. automatic, JPEG of ZIP voor kleurafbeeldingen) kiezen en de kwaliteit bepalen. In PDF-bestanden worden ook nog andere compressies gebruikt: LZW, RLE, CCITT Groep 3 en 4 (voor tekst, grafieken en monochrome afbeeldingen). De bestandsomvang van een PDF-bestand is doorgaans kleiner dan van een Word of Postcript-bestand. De Acrobat Reader of Exchange decomprimeren het bestand opnieuw.

Een PDF-bestand kan op drie wijzen worden weggeschreven: ongestructureerd, gestructureerd en getagd. Getagde PDF-bestanden zijn te verkiezen boven (on)gestructureerde. De tags maken het immers mogelijk dat ook andere applicaties paragrafen, tekstformattering, opsommingen en tabellen

[11] De optie 'ASCII-formaat' in de conversiesettings kan tot verwarring leiden. Hiermee wordt niet bedoeld dat de tekst als ASCII-karakters worden vastgelegd. ASCII-formaat wordt hier gebruikt als tegenovergestelde van binair bestand. Deze optie wordt het best gebruikt bij uitwisseling of om de bewerking van een PDF-bestand in een teksteditor mogelijk te maken.

herkennen en correct kunnen weergeven. Bij (on)gestructureerde PDF-bestanden is dit niet of veel minder het geval. Om een PDF-bestand op een getagde wijze te bewaren, moet de gebruiker enkel de optie 'Embed tags in PDF' aanvinken bij de conversiesettings (onder office). Deze tags kunnen niet echt met XML-tags worden vergeleken, maar eerder met HTML-tags. Adobe heeft deze optie in de eerste plaats voorzien voor het vastleggen van webpagina's in een PDF-bestand. De jongste twee PDF-specificaties voorzien in de mogelijkheid om bestanden te taggen of te structureren. Deze functionaliteit wordt echter pas Acrobat 5.0 geïntroduceerd.

In tegenstelling tot wat men zou vermoeden, kunnen PDF-bestanden relatief gemakkelijk worden aangepast. Er zijn diverse mogelijkheden waarop de inhoud van PDF-bestanden kan aangepast worden. Acrobat 4.0 bood enkel de mogelijkheid om PDF-documenten als PostScriptbestanden te exporteren. Versie 5.0 laat de gebruiker toe het PDF-bestand als RTF-bestand te bewaren zodat het in een tekstverwerkingsprogramma verder kan worden bewerkt. Wie een beetje handig is met een grafisch programma zoals Photoshop kan PDF-bestanden manipuleren. PDF-bestanden kunnen ook relatief gemakkelijk worden omgezet naar ASCII, RTF of HTML. Hiervoor worden aparte tools gebruikt[12].

Met Acrobat (Exchange) kan een PDF-bestand geannoteerd, gemarkeerd of gelinkt worden. Deze wijzigingen kunnen in het PDF-bestand worden opgeslagen.

Bij de omzetting van een tekstbestand naar PDF controleert men ook best of het PDF-bestand alle tekens bevat. Ervaring wijst uit dat bepaalde diakritische tekens niet mee opgenomen worden in het PDF-bestand en dat hun plaats gewoon wit is gebleven.

Vandaag de dag is PDF een veel gebruikt bestandsformaat en wordt het als een interessant archiveringsformaat naar voor geschoven. PDF heeft zijn populariteit te danken aan het feit dat men documenten in hun originele lay-out op een gemakkelijke manier resolutieloos kan vastleggen of uitwisselen. In Nederland wordt in de regeling geordende en toegankelijke staat archiefbescheiden PDF als archiveringsformaat voor tekst, afbeeldingen en CAD/CAM-tekeningen naar voor geschoven (art. 6).

Toch zijn er een aantal kanttekeningen. PDF is niet bruikbaar voor alle types tekstuele computerbestanden. PDF is in de eerste plaats bedoeld voor documenten die worden afgedrukt (brieven, rapporten, publicaties, enz) of die moeten uitgewisseld worden zonder dat de opmaak wijzigt. Hierdoor is de band met de papieren omgeving nog groot. PDF-bestanden hebben geen volwaardige ASCII-basis waardoor de afhankelijkheid van aangepaste software groot is. Het belangrijkste nadeel is ongetwijfeld de afhankelijkheid van één bepaalde producent. Adobe belooft wel een achterwaartse ondersteuning, maar het lijkt niet realistisch dat dit op lange termijn voor alle versies wordt volgehouden. Sommigen voorspellen PDF-bestanden een levensduur van 30 tot 50 jaar, maar daar is geen enkele garantie voor.

PDF-bestanden kunnen wellicht wel voldoen aan de vereisten voor bewaring op korte termijn, maar vanwege de producentgebondenheid lijkt het niet aangeraden om voor lange termijnbewaring digitale documenten als PDF-bestanden in het archief op te nemen. Er gaan wel stemmen op om PDF als ISO-standaard te laten vastliggen, maar dit lijkt nog veraf.

Referentie PDF: http://www.adobe.com; J.M. OCKERBLOOM, *Archiving and Preserving PDF Files*, in *RLG DigiNews*, febr. 2001, vol 1.; http://www.bvamyfra.fr/piproduc.htm;

[12] http://www.ra.informatik.uni-stuttgart.de/~gosho/pdftohtml/index.html; http://www.mosarca.com/bva-myfra/trap2gb.htm; http://www.cs.wisc.edu/~ghost/

B.2.5 Rich Text Format (RTF)

Het Rich Text Format is gecreëerd door Microsoft om de uitwisseling van tekstbestanden met opmaakgegevens tussen tekstverwerkingsprogramma's, en in het bijzonder tussen WordPerfect en Word, te vergemakkelijken. De kwaliteit van de omzetting van binaire tekstbestanden was immers in grote mate afhankelijk van de conversiefilters van de applicaties en had zelden een bevredigend resultaat. Om dit euvel te verhelpen, ontwierp Microsoft een gemeenschappelijk en open bestandsformaat. De specificatie van RTF is vrij beschikbaar op de Microsoft website. RTF 1.6 is de huidige versie.

Een RTF-bestand kan naast opgemaakte tekst ook afbeeldingen en grafieken bevatten. Een RTF-bestand bevat tekst, controlewoorden, controlesymbolen en groepen. Bij het maken van een RTF-bestand wordt de tekst duidelijk gescheiden van de code gegenereerd door de applicatieprogrammatuur. De codes worden vervangen door controlewoorden of commando's. De tekst en bijhorende controlewoorden en -symbolen worden samengebracht in groepen. In een groep worden opmaak en attributen van de bijhorende tekst beschreven. Om redenen van data uitwisseling kan een RTF-bestand enkel uit ASCII-karakters (7 bits) bestaan. RTF-bestanden bevatten verschillende lettertypes, voetnoten, annotaties, headers en footers, bladwijzers, hiërarchische kopteksten, secties, tabellen, enz. Een RTF-bestand kan enkel de low-level functies van een tekstverwerker zoals MS Word bewaren. Andere gegevens zoals macro's en opmaakstijlen gaan verloren. Gegevens over onder meer de gebruikte lettertypes, de gebruikte codetabel, de kleurentabel, de pagina-opmaak en het documentbeheer worden als headerinformatie opgeslagen.

RTF-bestanden kunnen afbeeldingen bevatten die met behulp van andere applicaties werden gemaakt. De binaire structuur wordt omgezet in een opeenvolging van cijfers en letters waarbij elk karakter 16 bits van de afbeelding bevat.

RTF-bestanden moet men in principe ook tussen verschillende platformen kunnen uitwisselen. In tegenstelling tot PDF (max. 255 karakters) kennen RTF-bestanden geen beperkingen in lijnlengte. RTF is hoofdlettergevoelig.

RTF-bestanden kenmerken zich door een grote bestandsomvang. Dezelfde opgemaakte tekst opgeslagen als RTF-bestand is gemiddeld 5 à 6 keer groter dan een MS Word-bestand.

Parallel met de uitbreidingen van de tekstverwerkingsprogramma's kent RTF een snelle evolutie. Er bestaan dus verschillende RTF-varianten. Een ander nadeel zijn de fouten die dikwijls met conversies gepaard gaan. Complexe RTF-bestanden met grafieken en afbeeldingen zijn soms corrupt.

RTF wordt veel gebruikt in desktop en officeapplicaties binnen Microsoft- en Appleomgeving. Er is geen enkele garantie op het vlak van duurzaamheid, zodat het gebruik van RTF voor archivering geen basisoptie is.

Referentie: http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnrtfspec/html/rtfspec.asp;

B.2.6 MS Word

Het *.doc-formaat van Microsoft Word is een applicatieformaat dat door zijn wijdverspreid gebruik een standaard is geworden. Het Wordformaat is niet ontworpen met het oog op uitwisseling, maar dit is in de praktijk wel mogelijk door het gebruik op grote schaal van het tekstverwerkingsprogramma MS Word. De recenste versies zijn Word 6.0, Word 97 en Word 2000. De applicatie Word is beschikbaar voor de verschillende courante platformen. Uitwisseling van hetzelfde Wordbestand tussen verschillende besturingssystemen is soms een huzarenstukje. Het doorsturen van het tekstbestand via het Internet kan hiervoor in de meeste gevallen een oplossing bieden.

De *.doc-bestanden zijn binaire bestanden. De tekst wordt bewaard als ASCII-karakters, maar de applicatie eigen codetekens worden aangegeven met behulp van hexadecimale of binaire tekens. De Wordbestanden zijn gebaseerd op OLE (Object Linking and Embedding). De tekst, de binaire data en afbeeldingen worden in afzonderlijke bitrepen opgeslagen die aan elkaar gekoppeld worden. De mainstream bevat een header en de tekst. De binaire data van de opgenomen objecten worden in de object streams bewaard.

Wordbestanden kunnen volledig opgemaakte teksten bevatten: tekst met opmaak, lettertypes, kleuren, headers en footers, voet-en eindnoten, indexen, bladwijzers, kruisverwijzingen, enz. In vergelijking met RTF-bestanden kunnen Wordbestanden ook high-level tekstverwerkingsfuncties bevatten zoals macro's en opmaakstijlen.

Of de inhoud van een *.doc-formaat correct wordt weergegeven, is afhankelijk van het hard- en softwareplatform waarop een Wordbestand wordt geopend. (bijv. geïnstalleerde lettertypes op het besturingssysteem).

Net zoals bij de andere meeste applicatieformaten is de achterwaartse compatibiliteit beperkt tot twee à drie generaties. De huidige Wordversie kan probleemloos bestanden inlezen die werden aangemaakt in Word 6.0 en Word 97. Het correct inlezen van nog oudere versies wordt al wat moeilijker. Wordbestanden zijn niet geschikt voor de bewaring op lange termijn. De ondersteuning is beperkt in tijd en volledig in handen van één producent. Overigens is het niet zeker dat er binnen afzienbare tijd geen andere tekstverwerker de standaard wordt. De bestanden aangemaakt met WordPerfect, de vorige standaard tekstverwerkingstoepassing, kunnen mits de nodige conversiefilters nog relatief gemakkelijk ingelezen worden. Voor Wordstarbestanden, de voorganger van WordPerfect, wordt dat al een beetje problematisch. Tekstverwerkingsbestanden hebben wel een ASCII-basis, maar bevatten heel veel binaire data die bij andere applicaties of hogere versies moeilijkheden opleveren.

Referentie: /

C. AUDIO-VISUELE BESTANDEN

C.1 Compressie en decompressie

De migratiestrategie is tot nu toe één van de meest toegepaste archiveringsmethoden voor digitale archiefdocumenten. Deze strategie houdt in dat digitale archiefdocumenten in een verouderd en niet meer ondersteund bestandsformaat naar de nieuwe standaard worden omgezet. De voorkeur gaat hierbij zoveel mogelijk uit naar een platformonafhankelijk formaat. Het comprimeren van computerbestanden gaat fundamenteel tegen deze doelstelling in. Door archiefdocumenten te comprimeren voegt men namelijk een bijkomende laag van software-afhankelijkheid toe. Immers, de gecomprimeerde bestanden zijn alleen raadpleegbaar wanneer men ze met de gepaste software kan decomprimeren. Na compressie is de output altijd 8-bits binaire data, ook al gaat het eigenlijk om gecomprimeerde ASCII-karakters. Compressie wordt best ook vermeden omdat het in veel gevallen in informatieverlies resulteert.

Compressie heeft als doel de originele bitstreams van een computerbestand (charstreams) door middel van algoritmes te herleiden tot kleinere bitstreams (codestreams). Dit biedt meerdere voordelen: sneller datatransport, kleinere bestandsomvang, sneller raadpleegbaar, minder zware computers nodig, er is minder bandwijdte vereist, enz. De bitstreams die het resultaat zijn van de compressiebewerking, vormen het nieuwe gecomprimeerde bestand. De verhouding in bestandsgrootte tussen het bronbestand en het bestand na compressie wordt aangeduid met de compressieratio.

De softwaretoepassing die men bij compressie gebruikt, wordt de *encoder* genoemd. De *decoder* voert de overeenstemmende decompressie uit zodat het bestand opnieuw raadpleegbaar is. Gecomprimeerde bestanden zijn binaire bestanden. De applicaties die als encoder functioneren zijn in de praktijk meestal ook in staat om het bestand te decomprimeren en te openen. Er zijn compressiemethoden die gebonden zijn aan één bepaald bestandsformaat, maar daarnaast zijn er ook compressiemethoden die bij verschillende bestandsformaten toepasbaar zijn. Een aantal van deze laatste groep compressiemethoden zou men als standaard kunnen beschouwen (bijv. LZW, RLE, JPEG, Huffman). MPEG en JPEG zijn compressietechnieken die zelfs als officiële standaard door ISO zijn vastgelegd. Anderzijds zijn er bestandsformaten waarbij verschillende compressiemethoden kunnen worden gebruikt (bijv. TIFF).

Compressie is gebaseerd op het vervangen van informatie die wordt herhaald en op het wegfilteren van informatie die door de mens niet wordt waargenomen. Er worden twee soorten compressie onderscheiden: lossless en lossy. Bij *lossless data compression* is het origineel bestand en het gedecomprimeerde bestand identiek. Na decompressie is de bitstream identiek aan de bitstream van voor de compressie. Deze methode wordt toegepast bij teksten, uitvoerbare computerbestanden en bepaalde afbeeldingscompressies. Er is geen informatie- of kwaliteitsverlies na compressie en decompressie. Het algoritme voor compressie en decompressie is hetzelfde, alleen worden de bewerkingen in omgekeerde volgorde uitgevoerd. Toch is lossless compression vanwege de bijkomende software-afhankelijkheid niet opportuun. Bovendien is de compressieratio bij lossless compression vrij beperkt. Dankzij de recentste technieken is het mogelijk om visually lossless (niet waarneembaar, wel dataverlies) of om lossless (geen dataverlies) te comprimeren.

Lossy data compression betekent dan dat er wel een verschil is tussen het origineel en het gedecomprimeerde computerbestand. Tijdens de compressie gaat er informatie verloren die bij decompressie niet wordt hersteld. Bij lossy compressie worden verschillende algoritmes voor compressie en decompressie gebruikt. Lossy data compression wordt veel toeegepast bij het bewaren

en uitwisselen van afbeeldingen, bewegend beeld en geluid. Met lossy compressie bereikt men een hogere compressieratio dan bij lossless compressie.

Bij de compressie van computerbestanden worden verschillende technieken gebruikt. Bij wavelet of fractale compressie worden de data van een computerbestand omgezet in wiskundige modellen. Bij wavelet compressie wordt de afbeelding opgesplitst in golven die de frequentie analyse van een functie weergeven. De vormen en patronen worden weergegeven door middel van wiskundige functies en formules. Met waveletcompressie kunnen in theorie compressieratio's tot 150:1 worden bereikt, maar in het algemeen schommelt de ratio tussen 100:1 en 15:1. Waveletcompressie wordt gebruikt bij foto's, geluid en video, 2D- en 3D-afbeeldingen en het digitaliseren van multimedia. Bij fractale compressie wordt de afbeelding beschreven met behulp van fractalen. Afbeeldingen bevatten ongeacht hun schaal fractalen die terugkerende vormen en patronen beschrijven. Met fractale compressie is op papier een ratio van 250:1 mogelijk. Normaal gezien varieert de compressieratio tussen 100:1 en 20:1. Het proces laat toe dat de afbeeldingen resolutie onafhankelijk worden opgeslagen. Hierdoor kan de afbeelding naar pixels worden omgezet zonder kwaliteitsverlies. Fractale compressie duurt opvallend langer dan wavelet compressie. Fractale en wavelet decompressie nemen ongeveer dezelfde tijd in beslag.

Een andere compressietechniek voor afbeeldingen is het bewaren van de data als een reeks (array) van pixels. De bekendste voorbeelden zijn de JPEG, de LZW en de Huffman-compressiemethode. JPEG behaalt een compressieratio tussen 10:1 en 20:1, LZW ongeveer 2:1. Volgens de officiële standaard kan JPEG lossless en lossy worden toegepast. In de praktijk wordt JPEG overwegend lossy toegepast. LZW-compressie heeft het statuut van defacto standaard. LZW-compressie is gebaseerd op de herhaling van strings in de data. Aangezien rasterafbeeldingen doorgaans heel wat stringherhalingen bevatten, is LZW bijgevolg dan ook vrij efficiënt. LZW is eveneens vrij snel in compressie en (de)compressie. LZW is lossless want alle informatie wordt bewaard. LZW werkt ook op bi-level afbeeldingen. LZW kent een ruimere verspreiding dan louter afbeeldingscompressie. LZW kan gebruikt worden om elk type digitaal bestand te comprimeren. Winzip is onder meer gebaseerd op LZW. Voor afbeeldingen in het algemeen is LZW doeltreffender dan Huffmancompressie. Huffmancompressie zet in tegenstelling tot LZW blokken data met een vaste lengte om in blokken met een variabele lengte.

Wavelet en fractale compressie slagen eigenlijk beter in hun opzet dan de compressiemethoden gebaseerd op het bewaren en bewerken van reeksen data. De wavelet en fractale compressiemethode hebben grote compressieratio's en verliezen minder informatie dan bijvoorbeeld JPEG. Het nadeel van deze compressiemethoden is echter dat ze nauwelijks gestandaardiseerd zijn en meestal eigendom zijn van één bepaalde producent. Vorig jaar kwam hier een wijziging in. De nieuwe JPEG2000-standaard is een toepassing van waveletcompressie.

De algemene aanbeveling bij digitale archivering blijf echter ongecomprimeerde opname van de digitale archiefstukken in het digitaal depot. Dit geldt zeker voor de digitale moederkopieën. Voor on line terbeschikkingstelling zal het gebruik van bestandsformaten met compressie meer aangewezen zijn. Deze aanbeveling is echter niet altijd gemakkelijk te realiseren. Vooral bij digitale bestanden met bewegend beeld en geluid is compressie van de moederkopieën moeilijk te vermijden. Zonder compressie swingt de omvang van audio-visuele bestanden de pan uit, wat tot moeilijkheden bij uitwisseling of bewerking zou leiden. Eén minuut ongecomprimeerde digitale videobeelden (30 frames/seconde) neemt 1,66 gigabyte in beslag. Bij de keuze van een geschikt archiveringsformaat voor digitale televistoe- of videobeelden moet er bijgevolg gelet worden op het gebruik van gestandaardiseerde compressiemethoden en een zo klein mogelijk kwaliteitsverlies.

Bij digitale afbeeldingen en audiosignalen is het compressieloos archiveren van de digitale moederkopieën gemakkelijker haalbaar. Een compressieloze hoge resolutie TIFF-afbeelding neemt ongeveer 4 à 5 megabytes in beslag, wat geen noemenswaardige opslagproblemen oplevert. Voor de verspreiding via Internet kunnen er lage resolutie en gecomprimeerde kopieën (bijv. JPEG) worden gemaakt. Voor een compressieloos stereogeluidssignaal van 1 minuut is ongeveer 10 MB nodig. Voor uitwisseling via een netwerk kan een gecomprimeerde versie gemaakt worden. Een bijzondere compressietoepassing is streaming[13].

Een laatste aandachtspunt bij het evalueren van compressiemethoden is de auteursrechtelijke bescherming van de (de)compressiealgoritmes. Bepaalde compressiemethoden zijn immers auteursrechtelijk beschermd zodat men over een gebruikerslicensie dient te beschikken[14].

C.2 Audio-visuele bestanden

C.2.1 Officiële standaarden

C.2.1.1 MPEG-Video

MPEG is de benaming van een ISO-werkgroep die in 1988 werd samengesteld en staat voor Moving Pictures Experts Group. MPEG werd in het leven geroepen om compressiestandaarden voor de digitale opslag van video, audio en een combinatie van beide te creëren. Met de benaming MPEG wordt de officiële standaardfamilie voor audio-visuele computerbestanden aangeduid. De MPEG-standaarden zijn open. Ondertussen bestaan er al verschillende MPEG-versies: MPEG-1, MPEG-2, MPEG-4, MPEG-7 en MPEG-21. MPEG wordt gebruikt voor computerbestanden die audio en/of visuele informatie bevatten. De standaarden van de MPEG-familie worden volop geïmplementeerd in commerciële toepassingen. De extensie voor MPEG-Video is *.mpeg of *.mpg. De videocompressie is deels gebaseerd op JPEG. Voor MPEG-Audio wordt de extensie *.mp gehanteerd. De MPEG-Audio standaard wordt besproken op p. **Fout! Bladwijzer niet gedefinieerd.**.

De MPEG-Video familie biedt standaarden aan voor bewegende digitale beelden met/zonder geluid. De verschillende MPEG-Video groepen leveren verschillende kwaliteiten en overdrachtsnelheden. MPEG-bestanden kunnen bekeken worden in diverse applicaties (Quicktime Player, Windows Media Player, enz.).

MPEG-1 levert bewegende beelden aan VHS-videorecorder outputkwaliteit aan ongeveer 1,2 a 1,5 Mbps. MPEG-1 was vooral ontworpen voor CD-I en Video-CD. De courante toepassingen van MPEG-1 leveren een videoresolutie van 352 pixels op 240 lijnen aan 30 frames per seconde (“Low Level”)[15].

---

[13] Streaming: een technologie die het mogelijk maakt dat audio- en video-bestanden worden geopend naar mate ze worden ontvangen. Wat binnenkomt, wordt onmiddellijk getoond waardoor men niet hoeft te wachten tot dat het volledige bestand is gedownload.

[14] http://www.unisys.com/unisys/lzw/

[15] MPEG-1: *ISO/IEC 11172 (1992): Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s*

MPEG-2[16] werd ontworpen voor digitale televisie aan uitzendkwaliteit (Standard Definition Television) met een transmissiesnelheid tussen 4 en 6 Mbps. MPEG-2 wordt hoofdzakelijk door de (digitale) televisie- en DVD-industrie gebruikt[17]. MPEG-2 past onder andere interlaced videosignalen toe.

MPEG-4 is de officiële standaard voor de bundeling van multimediabestanden, interactieve afbeeldingen en digitale TV binnen netwerktoepassingen. MPEG-4 werkt vooral op basis van lage sample-rates en lage data-encoding en splitst de onderdelen van een multimediatoepassing als afzonderlijke objecten op[18]. Als basis voor MPEG-4 werd het Quicktimeformaat gebruikt[19].

MPEG-7 ("Multimedia Content Description Interface") is niet bedoeld om MPEG-4 te vervangen, maar vult MPEG-4 aan. Terwijl MPEG-4 wordt gebruikt voor de weergave van een bepaalde inhoud, geeft MPEG-7 aan hoe de multimedia inhoud wordt beschreven, opgezocht en beheerd. MPEG-7 biedt specifieke en gestandaardiseerde tools aan voor de beschrijving van metadata als gestructureerde informatie. Het toepassen van MPEG-7 moet de uitwisseling van MPEG-4-toepassingen vergemakkelijken of mogelijk maken. MPEG-7 maakt ondermeer gebruik van XML, XML Schema, Dublin Core Metadata, enz.

MPEG-21 is momenteel in voorbereiding[20].

Referentie: http://www.iso.ch; http://mpeg.telecomitalialab.com/; http://www.mpeg.org

C.2.2 Defacto standaarden

C.2.2.1AVI: Audio Video Interleave

AVI is Microsofts RIFF-toepassing (Resource Information File Format) voor de opslag van video met bijhorend geluid. De extensie is *.AVI. Hun MIME-type kan op verschillende manieren worden aangeduid: video/avi, video/msvideo of video/x-msvideo.

De audio- en videodata in een AVI-bestand kunnen op basis van verschillende compressies en codecs worden opgeslagen (video codecs: o.a. MPEG, Microsoft Video, Intel Indeo, Cinepak Codec, VDOwave, Motion JPEG – audio codecs: o.a. PCM, MP3, ADPCM). De frames hoeven niet gecomprimeerd te worden. In dit geval wordt de codec aangeduid met DIB, RGB of RAW. Het aantal frames per seconde en de sample-rate is instelbaar en aanpasbaar. Het aantal frames per seconde (doorgaans 30) kan verminderd worden om de bestandsgrootte te verkleinen. Hierbij gaan frames verloren en kan de slow motion functionaliteit wegvallen.

---

16 MPEG-3 was bedoeld voor High Definition Televisie toepassingen. MPEG-2 bleek echter aan behoeften hiervoor te voldoen zodat dit standaardiseringsinitiatief bij MPEG-2 werd ondergebracht.

17 MPEG-2: *ISO/IEC 13818 (1994): Information Technology--Generic Coding of Moving Pictures and Associated Audio.*

18 MPEG-4: *ISO/IEC 14496 (1998): Information technology -- Coding of audio-visual objects*

19 http://www.apple.com/pr/library/1998/feb/11iso.html

20 MPEG-21: *ISO/IEC TR 21000-1(2001): Information technology -- Multimedia framework (MPEG-21) -- Part 1: Vision, Technologies and Strategy*

AVI-bestanden zijn in de eerste plaats bedoeld om afgespeeld te worden op Windowsplatformen. AVI-bestanden zijn minder gebruikt en minder gemakkelijk uit te wisselen dan Quicktimebestanden. AVI wordt voor verspreiding via het Internet veelal omgezet naar het ASF-videostreamingformaat (eerst Active Stream Format, nu Advanced Streaming Format) van Microsoft. Er is ook een OpenDML AVI M-JPEG File Format gemaakt. Dit is een AVI-compatibel bestandsformaat voor professionele video.

C.2.2.2Quicktime

Het bestandsformaat Quicktime is ontworpen door Apple Computer. Quicktimebestanden zijn inmiddels ook afspeelbaar op Unix- en Windowscomputers. Apple/MacIntosh schuift Quicktime bijgevolg als een geschikt uitwisselingsformaat naar voor. Een Quicktimebestand kan bijna om het even welk type digitaal bestand bevatten: audio, video, 3D, animatie, afbeeldingen en virtuele realiteit. Quicktime is dan ook een echt multimediaformaat. De recenste specificatie van het bestandsformaat dateert van maart 2001. De specificatie is vrij beschikbaar. Quicktimebestanden hebben de extensie *.mov, *.moov of *.qt. Hun MIME-type is “video/quicktime”. Tenzij anders bepaald worden de data in big-endian volgorde opgeslagen. Een groot deel van de MPEG-4 standaard is op Quicktime gebaseerd.

De structuur van een Quicktimebestand is een hiërarchische indeling van geneste atomen. Er zijn basisatomen en optionele atomen. De volgorde van de atomen is in principe vrij. De meeste Quicktimebestanden zijn gecomprimeerd. Er zijn verschillende compressietechnieken of codecs bij Quicktimebestanden toepasbaar. (beelden: o.a. JPEG, Cinepak, Apple Video, Kodak Photo CD en MPEG – geluid: o.a. MACE, µ-law, A-law, MP3 en ADPCM).

Samen met MPEG is Quicktime één van de meest voorkomende bestandsformaten voor bewegend beeld en geluid op het Internet. Er kan een heel gamma bestandsformaten naar Quicktime worden omgezet. In de Quicktimespecificatie is aandacht besteed aan de metadata die in deze bestandsformaten zijn ingekapseld. Voor de metadata in AVI, MP3, WAV, FlashPix, (animated) GIF, JFIF/JPEG, Photoshop en TIFF wordt duidelijk aangegeven naar welke Quicktime velden deze metadata worden gemapt. Anderzijds kan de inhoud van een Quicktimebestand naar verschillende bestandsformaten worden omgezet.

Quicktime ondersteunt ook streaming audio en streaming video.

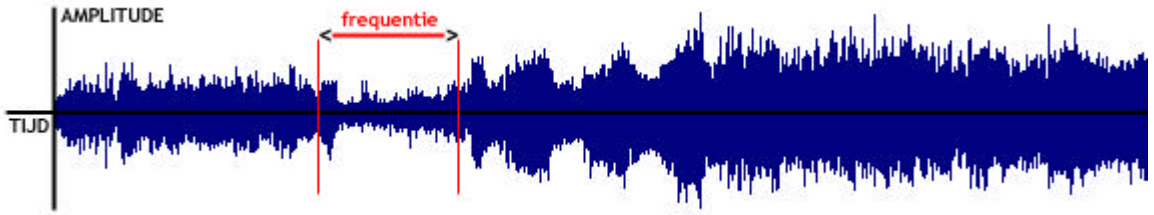
Referentie: http://developer.apple.com/techpubs/quicktime/qtdevdocs/QTFF/qtff.html

C.3 Audio bestanden

C.3.1 Van analoog naar digitaal: het digitaliseringsproces

Geluid plant zich als analoge golven of trillingen in de ruimte voort. Een microfoon zet de geluidsgolf in een stroom van wisselende elektrische spanningen om. Bij analoge opslag zet een transductor de

elektrische spanningen om naar fysieke veranderingen in de groef van een LP of naar een magnetische ordening op een magnetische band. Digitalisering heeft als doel de analoge geluidsgolf in binaire data om te zetten. Het digitaliseringsproces van audiosignalen verloopt in twee stappen. Ten eerste meet een A/D convertor de amplitude van de geluidsgolf. De A/D convertor meet niet de volledige golf maar voert met een bepaalde frequentie metingen uit langs de geluidsgolf. De frequentie wordt de sampling-rate genoemd en wordt uitgedrukt in (kilo)herz. (Kilo)herz geeft het aantal metingen per seconde aan. Een waarde van 44,1 Khz. geeft aan dat per seconde de geluidssterkte 44100 keer wordt gemeten. Hoe hoger de frequentie, hoe kleiner de intervallen en hoe preciezer de digitale geluidsgolf de oorspronkelijke geluidsgolf weergeeft.



Afbeelding 2: Een digitaliseerde geluidsgolf (mono). Het analoog audiosignaal wordt gereconstrueerd door de meetresultaten met elkaar te verbinden. Bij een stereo geluidsgolf worden twee kanalen gebruikt. Het linker kanaal gaat doorgaans het rechtse vooraf.

Tijdens de tweede stap zet een convertor de meetresultaten om in numerieke binaire gegevens. Voor de digitale opslag van deze gegevens worden bits gebruikt. Het aantal bits varieert doorgaans tussen 8, 16 en 24 bits en wordt aangeduid met de term sample-resolutie. Hoe meer bits er worden gebruikt, hoe accurater het meetresultaat digitaal wordt opgeslagen. Men kan dit het best vergelijken met de gedetailleerdheid van de schaal waarop de digitale amplitude wordt uitgetekend. Op een schaal van 0 tot 16,7 miljoen (24 bits) kan de geluidssterkte exacter worden uitgetekend dan op een schaal van 0 tot 255 (8 bits). Een grotere sample-resolutie verhoogt de dynamic range[21] en beperkt het achtergrondgeluid. Een hogere sample-resolutie resulteert vooral in de betere omzetting van lage toonsignalen. Samen met het aantal kanalen[22] (1 voor mono, 2 voor stereo) zijn de sample- rate en de sample-resolutie de voornaamste parameters in het digitaliseringsproces.

Het is duidelijk dat bij digitalisering informatie van de oorspronkelijke analoge geluidsgolf verloren gaat. Het informatieverlies is het kleinst bij een zo hoog mogelijke sampling-rate en bij het gebruik van zoveel mogelijk bits bij de digitale opslag van de metingen. Het gevolg is dan natuurlijk een grote omvang van het digitale bestand. Voor een gedigitaliseerd (16 bits) stereosignaal (2 kanalen) van één minuut (60 seconden) aan CD-kwaliteit (44100 Khz/sec) is ongeveer 10 MB vereist.

Bij de digitalisering van analoge geluidsgolven wordt doorgaans het PCM digitaal schema gebruikt. PCM staat voor Pulse Code Modulation en is het standaard binair formaat voor ongecomprimeerde gesamplede audiosignalen. Na de digitalisering wordt het geluid in PCM de digitale moederkopie. De PCM-kwaliteit is vooral afhankelijk van de sampling-rate en het aantal bits

[21] Dynamic range: het verschil tussen het hoogste en het laagste signaal. Bij 16 bits bedraagt de dynamic range 96 dB.

[22] 2 kanalen: links, rechts
3 kanalen: links, rechts, midden
4 kanalen: links, rechts, midden, surround
6 kanalen: links midden, links, midden, rechts, rechts midden, surround

voor de numerieke code. Deze twee factoren worden doorgaans met twee cijfers weergegeven: bijv. 24/96 (24 bit/96 Khz).

De sample-rate en sample-resolutie die bij digitalisering best wordt gebruikt is afhankelijk van de kwaliteit de bron. Wanneer digitalisering als archiveringsstrategie voor analoog geluid wordt gebruikt, dan moet het gedigitaliseerde signaal een zo getrouw mogelijke weergave van het analoog signaal zijn[23]. Algemeen wordt aangeraden om aan een maximaal mogelijke sampling-rate en sampling-resolutie te digitaliseren[24]. Nadeel van deze optie is de grote bestandsomvang. Bovendien heeft dat ook weinig zin bij een analoge bron van lage kwaliteit. Vaste sample-rates en sample-resoluties hanteren heeft dan ook niet veel nut. In de praktijk worden echter meestal vaste parameters toegepast. Voor gesproken woord is 8000 Herz en 8 bits veel gebruikt, terwijl voor muziek een hogere sample-rate en sample-resolutie nodig is. De *Audio Engineering Society* legde in haar standaard een sample-rate van 48 Khz. vast, maar liet nog ruimte voor 44,1 Khz. (voor commerciële toepassingen), 32 Khz. (voor uitwisselingstoepassingen) en 96 Khz. (voor toepassingen met grote bandwijdte)[25].

Bij de opslag van digitale moederkopieën is comprimering best te vermijden. Dit vergroot niet alleen de platformafhankelijkheid van de digitale bestanden, maar wordt best ook afgewezen vanwege het informatieverlies waarmee lossy compressie gepaard gaat[26].

Het digitale geluid kan als een audiotrack op een audiodrager worden opgeslagen of kan als een binaire computerbestand worden opgeslagen. Bij de keuze van een drager voor audiotracks moet men wel rekening houden met eventuele beperkingen op sample-rates en sample-resoluties. Een audio-CD is interessant vanwege de grote compatibiliteit en het vermijden van software-afhankelijkheid. PCM is trouwens het standaardformaat waarin digitaal geluid op een audioCD's of een DVD-audio wordt opgeslagen. Nadeel van audioCD's is echter dat ze standaard 16 bits/44,1 Khz gebruiken. Deze waarden zijn zo gekozen vanwege de perceptuele beperkingen van het menselijk oor. De bijkomende informatie die bij een hogere sample-rate of sample-resolutie wordt opgeslagen, is voor niet-specialisten immers niet of nauwelijks waarneembaar. Bij de opslag van een 24/96 digitaal signaal op een audioCD gaat dus opnieuw informatie verloren. Een Digitale Audio Tape (DAT) kan sampling-rates van 32 / 44,1 / 48 Khz bevatten, maar de DVD-audio biedt de beste mogelijkheid om een geluid in PCM zonder kwaliteitsverlies als audiotrack te bewaren. Een alternatief voor PCM is Direct Stream Digital (DSD). DSD past lossless compressie toe. DSD ondersteunt 100 kHz frequentie en 120 DB dynamic range.

Tabel 2: Vergelijking tussen DVD-Audio en CD-Audio.

PARAMETER	DVD-AUDIO	CD-AUDIO
coding	PCM	PCM
aantal kanalen	max. 6 kanalen	2 kanalen
frequency respons	0 tot 96 Khz	5 tot 20 Khz
dynamisch bereik	144 db	96 db

[23] INTERNATIONAL ASSOCIATION OF SOUND AND AUDIOVISUAL ARCHIVES, *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy (versie, september 2001):* art. 9.

[24] Bijv.: J. FLEMING, *Forward into the past - protecting our musical heritage*, in: *Journal of the Audio Engineering Society*, vol. 49, nr. 7/8, juli/augustus 2001, p. 677.

[25] *AES recommended practice for professional digital audio — Preferred sampling frequencies for applications employing pulse-code modulation*, 1998.

[26] INTERNATIONAL ASSOCIATION OF SOUND AND AUDIOVISUAL ARCHIVES, *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy (versie, september 2001):* art. 10.

sample-rate bij 2 kanalen	44,1 – 88,2 – 176,4 Khz 48 – 96 – 192 Khz	44,1 Khz
sample-rate multi-channel	44,1 – 88,2 Khz. 48 – 96 Khz.	/
sample-resolutie	12, 16, 20 of 24 bits	16 bits
max. data rate[27]	9,6 mbps	1,4112 mbps

De betere geluidskwaliteit op een DVD-audio is het resultaat van de veel grotere opslagcapaciteit van de DVD-schijf in vergelijking met een CD-schijf. Om die opslagcapaciteit nog te vergroten, heeft het DVD-Forum een alternatief formaat voor PCM ontwikkeld: Meridian Lossless Packing (MLP). Dit is een lossless compressieschema dat in vergelijking met ongecomprimeerde PCM een compressieratio van 2:1 bereikt. Bij MLP is de compressie niet gebaseerd op perceptuele of andere weglatingen zodat er geen kwaliteitsverlies optreedt. Bij Dolby Digital en DTS (onder meer gebruikt voor de opslag van geluid op DVD-video) is dit wel het geval. Eén zijde van een DVD-audio kan in MLP bijgevolg twee uren geluid in 6 kanalen met een 24/96 kwaliteit bevatten.

Eens een analoog signaal in een digitaal is omgezet, is ook opslag als een binair computerbestand mogelijk. De verschillende bestandsformaten voor geluid zijn onder meer: AU, AIFF, MP3, PCM, QT, RM, WAV en WMA. Er wordt een onderscheid gemaakt tussen de ruwe en zelfbeschrijvende audiobestanden. PCM wordt als een ruw audiobestand omschreven omdat het geen header heeft en enkel de audiodata in PCM bevat. De zelfbeschrijvende audioformaten bestaat uit twee delen: de header en de audio data. In de header wordt informatie over het geluidsbestand (de lengte, de sample-rate, de sampling-resolutie en eventueel de compressie) opgeslagen. De audiogegevens zelf zijn volgens een welbepaalde structuur en codering (codec) in het gedeelte met de audio data opgeslagen. PCM is één mogelijkheid en wordt ondermeer in AIFF en WAV gebruikt. PCM is compressieloos. AIFF- en WAV-bestanden hebben bijgevolg een grote bestandsomvang. Andere codec-mogelijkheden zijn µ-law (AU), A-law, MPEG-1 Audio Layer 3 en aantal producentgebonden codecs (Windows Media Audio: Microsoft; RealMedia: Real Networks; QuickTime: Apple).

De binaire computerbestanden met hoge sampling-rates kunnen in principe op om het even welke drager worden opgeslagen. CD-ROM's kunnen 40 minuten 24bits/48Khz of 20 minuten 24bits/96Khz geluid bevatten.

Voor terbeschikkingstelling via het Internet worden geluidsbestanden veelal ook omgezet naar streaming audiobestanden. Bij het bewaren als streaming bestanden wordt altijd compressie toegepast, wat de kwaliteit niet ten goede komt. Digitale archiefdocumenten kunnen bijgevolg niet als streamingbestanden in het digitaal depot worden opgenomen. Voorbeelden van streamingbestanden zijn Advanced Streaming Format (ASF, audio en video), streaming MP3 (audio),

Voor het beluisteren van het gedigitaliseerd signaal is vervolgens een D/A convertor nodig die de digitale data opnieuw naar een analoge geluidsgolf omzet. Een D/A convertor is doorgaans in de afspeelapparatuur ingebouwd (bijv. CD-spelers).

[27] Max. data rate: aantal miljoen bits per seconde die maximaal verwerkbaar zijn

C.3.2 Officiële standaarden

C.3.2.1 MPEG-Audio

MPEG-1 Audio onderscheidt drie verschillende compressieschema's met een eigen performantie (layer 1: 4:1 PASCcompressie die onder andere in Digital Compact Cassettes wordt gebruikt; layer 2: 6:1 tot 8:1 MUSICAMcompressie; layer 3: 10:1 tot 12:1 compressie). Door de toepassing van de MP3 datareductieschema's kan de bestandsomvang zodanig afnemen zonder dat de sample-rate moet verminderd worden. De audiobestanden gecomprimeerd op basis van MPEG-1 Audio layer 2 worden *.mp2-bestanden genoemd. Het populaire *.mp3-bestand is het bestandsformaat waarbij de MPEG-1 Audio layer 3 comprimering werd gebruikt. MP3 levert de hoogste kwaliteit en compressieratio binnen de MPEG-1 audiostandaard, wat zijn populariteit op het Internet verklaart. Binnen de MPEG-1 familie is MP3 ook de meest complexe. MP3 gebruikt 32 / 44,1 / 48 Khz sample-frequenties. De bit-rate van MP3-bestanden is niet vast en kan variëren van 32 kbit/sec tot 320 kbit/sec voor een stereosignaal. De MP3-bestanden kunnen opgeslagen worden met een vaste (CBR) of een variable (VBR) bitrate. MP3 is toepasbaar op mono, stereo, twee onafhankelijke kanalen op en joint stereo[28].

Aan een MP3-bestand kan achteraan een ID3-tag worden toegevoegd. In deze ID3-tag kan men volgende gegevens toevoegen: titel, artiest, album, genre, jaar en commentaar. Deze gegevens worden op het tijdstip van de encoding toegevoegd of men kan ze achteraf met een ID3-editor aan het MP3-bestand toevoegen. De ID3-tag wordt geïdentificeerd door het veld 'TAG' en bevat de metadata in de vorm van ASCII-karakters.

MPEG-2 Audio is een uitbreiding van MPEG-1 Audio en voegt een codering aan lagere sample-rates (16 / 22,05 / 24 Khz) toe. MPEG-2 Audio past dezelfde compressieschema's als MPEG-1 toe. Bij tests stelde men vast dat het gebruik van andere coderingsalgoritmes grotere compressieratio's opleverde. Dit onderzoek leidde in 1997 tot de MPEG-2 Advanced Audio Coding (AAC) standaard[29]. AAC is bedoeld als opvolger van MP3. In vergelijking met MP3 heeft AAC een grotere compressieratio met minder kwaliteitsverlies. Er is wel geen compatibiliteit tussen MP3 en AAC. Het Fraunhofer instituut heeft een auteursrechtelijke beschermde uitbreiding op de MPEG-2 audio met 8 / 11,05 / 12 Khz. Deze uitbreiding wordt ook wel eens MPEG-2,5 genoemd.

C.3.2.2PCM: Pulse Code Modulation

PCM is de wijze waarop ongecomprimeerde digitale audiosignalen gewoonlijk worden opgeslagen en overgebracht. PCM wordt toegepast bij audioCD's, audioDVD's en digitale audiotapes (DAT's). Bestandsformaten zoals WAV en AIFF gebruiken de PCM-codec. De meeste audiocomputertoepassingen kunnen PCM inlezen. De bits van een PCM-bestand staan de I's rechtstreeks voor een pulse en de 0's voor de afwezigheid van een pulse. PCM wordt wel eens vergeleken met ASCII voor tekstbestanden. Ongecomprimeerde PCM wordt ook wel eens LPCM

[28] Twee onafhankelijke kanalen: vb. 1 taal per kanaal
Joint stereo: efficiënte combinatie van twee het linker en rechter kanaal

[29] ISO/IEC 13818-7:1997 Information technology -- Generic coding of moving pictures and associated audio information -- Part 7: Advanced Audio Coding (AAC)

(Lineaire PCM) of raw PCM genoemd. PCM is vastgelegd in de ITU-Recommendation G.711. Een veel gebruikte PCM-kwaliteit is 24 bits/48 Khz. (komt ongeveer overeen met 1 Gigabyte/uur).

Een gedigitaliseerde geluidsgolf kan als een *.pcm/*.raw-bestand worden opgeslagen. Een dergelijk bestand wordt als een ruw bestand of een dump van de audiodata beschouwd. Een pcm-bestand heeft in tegenstelling tot zelfbeschrijvende audiobestanden (bijv. AU, AIFF en WAV) geen file header waarin technische gegevens over de audiodata zijn opgeslagen. Bij opening van het bestand dient de gebruiker bijgevolg zelf de sample-rate, de sample-resolutie en het aantal kanalen op te geven. Om de eventuele problemen te vermijden kan men de belangrijkste (header)gegevens in een *.DAT-bestand bijhouden. Een DAT-bestand bevat dan bijvoorbeeld: 44100, 16, 2, PCM, Intel. Deze gegevens staan voor 44100 herz, 16 bits, 2 kanalen, PCM-code en Intel-encoding. De audiodata in een PCM kunnen zowel in big-endian als in little-endian volgorde worden opgeslagen.

Er bestaan varianten op PCM met de bedoeling om via compressie de hoeveelheid audio data te reduceren. DPCM (Differential Pulse Code Modulation) is een eenvoudig lossy compressietoepassing waarbij enkel het verschil tussen twee opeenvolgende samples wordt bewaard. Ongeacht de oorspronkelijke sample-resolutie van het bronbestand gebruikt DPCM altijd 4 bits. De compressieratio varieert dus al naargelang het bronbestand. ADPCM (Adaptive Differential Pulse Code Modulation) is gebaseerd op DPCM waarbij de sample-resolutie wordt aangepast aan de complexiteit van het audiosignaal. ADPCM gebruikt 16 / 24 / 32 / 40 bits voor de opslag van de binaire amplitudewaarden. Er bestaan een aantal varianten op ADPCM. ADPCM is vastgelegd in ITU-Recommendation G.726 en G.727 en in de DVI-standaard van de Interactive Multimedia Asscociation. ADPCM bestaat ook in producentgebonden versies (Microsoft, Creative Labs, enz.).

C.3.3 Defacto standaarden

C.3.3.1WAV

Het WAV-formaat werd ontwikkeld door Microsoft en IBM en wordt als defacto standaard voor geluidsbestanden op Windowscomputers gebruikt. De volgorde van de data is little-endian (Intel).

WAV is Microsofts toepassing RIFF voor de opslag van audiobestanden. De WAV-bestanden zijn doorgaans niets meer dan een RIFF-header die wordt gevolgd door verschillende chunks. De RIFF-header bestaat uit de letters RIFF (als ID-chunk), de chunksize en de letters WAVE waarmee het formaat wordt aangeduid. Na de RIFF-header volgen de twee basischunks van het WAV-formaat. De format-subchunk (ID-subchunk 'fmt') identificeert de audiodata (ID-subchunk 'data') en bevat informatie over het audioformaat en eventuele compressie (1 = compressieloos PCM), het aantal kanalen (1 = mono, 2 = stereo, enz.)., de sample-rate en de bit-rate (=sample-rate x aantal kanalen). Bij een gecomprimeerd WAV-bestand worden bijkomende velden aan de formatchunk toegevoegd die bij de decompressie worden gebruikt. Na de ASCII-karakters 'DATA' volgen de vermelding van de resterende chunkgrootte en de audiodata. De audiodata zelf zijn meestal gecodeerd op basis van PCM. WAV wordt bijgevolg soms als Windows PCM aangeduid. Samen met de wijdverspreidheid van Windowscomputers heeft dit voor gevolg dat ongecomprimeerde WAV-bestanden relatief gemakkelijk uitwisselbaar zijn. De PCM-codec heeft wel een grote bestandsomvang voor gevolg. Voor 1 minuut geluid met CD-kwaliteit (16/44,1) is ongeveer 10 megabytes nodig. WAV-bestanden

worden dan ook niet veel gebruikt binnen netwerktoepassingen. Voor uitwisseling over het Internet worden WAV-bestanden dan ook doorgaans naar MP3 omgezet of wordt binnen WAV de MP3-codec gebruikt. Alle andere chunks in het WAV-formaat zijn optioneel (cue, playlist, associated data, instrument, enz.).

In de plaats van PCM kunnen ook andere codecs worden gebruikt om de data in een WAV-bestand op te slaan: A-law, µ-law, ADPCM, MP3, enz. Een ongecomprimeerd WAV-bestand (PCM-codec) dat wordt omgezet naar een WAV-bestand met MP3-codec wordt 20:1 kleiner.

In een WAV-bestand kan een geluidssignaal in verschillende sample-rates (van 6000 tot 192000 Hz) en in verschillende sample-resoluties (van 8 tot 32 bits) worden opgeslagen. In een WAV-bestand kan ook gebruikersinformatie worden ingebed. Deze gegevens worden opgeslagen in labeled textchunk. De standaard RIFF-header voorziet volgende metadatavelden: titel, artiest, album, genre, trefwoorden, digitale bron, medium, ingenieurs, digitizer, leverancier, copyright, software en creatiedatum.

WAV kent een grote toepassing op zowel personal computers als in professionele opname-apparatuur. Het EBU (Euoropean Broadcast Union) heeft WAV verder ontwikkeld tot BWF (Broadcast Wave Format) zodat uitwisseling tussen de verschillende Europese radiostations mogelijk is.

Referentie: /

C.3.3.2AU / SND

AU (Access Unit) is het audioformaat dat is ontwikkeld door Sun Microsystems en NeXt. AU- en SND-geluidsbestanden hebben intern dezelfde structuur. Het AU-formaat gebruikt doorgaans µ-law als codec, maar A-law en (AD)PCM zijn eveneens mogelijk. De codec µ-law slaat zijn data in 8 bits op, maar in tegenstelling tot andere bestandsformaten past µ-law logaritmische ipv lineaire encoding toe. Hierdoor wordt een dynamisch bereik gehaald dat het equivalent is van 12-bits opslag. Nadeel is wel dat bestanden met logaritmische encoding meer ruis bevatten dan bestanden met lineaire encoding.

AU-bestanden zijn in grote mate platformonafhankelijk en kunnen ook door andere besturingssystemen dan Unix, Linux of Solaris worden ingelezen. Veel Windowstoepassingen kunnen *.AU-bestanden openen. AU wordt dan ook regelmatig gebruikt als uitwisselingsformaat over het Internet. AU-bestanden kunnen gecomprimeerd worden bewaard, maar de toepassing van compressie maakt uitwisseling moeilijker vanwege de bijkomende compatibiliteitsproblemen met andere platformen.

AU- en SND-bestanden ondersteunen verschillende sample-rates en meerdere kanalen. Een AU-bestand is samengesteld uit drie blokken: de header, het annotatieblok en de audiodata.

Referentie: /

C.3.3.3AIFF: Audio Interchange File Format

Het AIFF-geluidsformaat wordt in de eerste plaats door Apple/MacIntosh-computers gebruikt. Ook in professionele opname-omgevingen wordt AIFF frequent gebruikt. AIFF-bestanden hebben de extensies *.aif of *.aiff

AIFF is ontworpen met de bedoeling om de uitwisseling van geluidsbestanden tussen verschillende platformen mogelijk te maken. AIFF-bestanden kunnen verschillende sample-rates en bitdiepten ondersteunen. AIFF laat hoge digitale kwaliteit toe. AIFF-bestanden kunnen gecomprimeerd of ongecomprimeerd worden opgeslagen. De gecomprimeerde AIFF-bestanden worden AIFF-C of AIFC genoemd.

AIFF gebruiken de PCM-codec voor de opslag van de geluidsdata. Het AIFF-bestand is net zoals WAV samengesteld uit chunks. Er zijn twee basischunks die in elk AIFF-bestand voorkomen: de common chunk ("COMM") en de sound data chunk ("SSND"). De common chunk kan met een fileheader worden vergeleken. Hierin worden de parameters van de geluidsgolf beschreven: lengte, sample-rate, sample-resolutie, aantal kanalen, enz.. In de sound data chunk worden de eigenlijke geluidsdata bijgehouden. Alle andere chunks zijn optioneel (marker, instrument, MIDI data, audio recording, comments, text chunks, enz). In de text chunk is plaats voorzien voor de titel, uitvoerdersnaam, copyright en annotaties. Gebruikers kunnen in principe datachunks voor eigen gebruik toevoegen. Alle data wordt in big endian formaat opgeslagen. Alle AIFF-compatibele applicaties moeten op zijn minst de twee basischunks kunnen inlezen. Optionele chunks kunnen genegeerd worden. Er kunnen ook chunks worden toegevoegd die eigen zijn aan een bepaalde applicatie of gebruik, maar deze chunks worden eveneens genegeerd door computerprogramma's die ze niet ondersteunen.

Referentie: *Audio Interchange File Format (AIFF): A Standard for Samples Sound Files, Version 1.2*

C.3 Afbeeldingen

Digitale afbeeldingen worden doorgaans in twee groepen verdeeld: bitmap/rasterafbeeldingen en vectorafbeelden.

De bitmap- of rasterafbeeldingen worden opgeslagen als een verzameling pixels gerangschikt in rijen en kolommen. Elk punt van de afbeelding (een pixel) kan met een tabelcel worden vergeleken. De bitmap bevat de afbakening van de afbeeldingsruimte en de kleuren van de pixels binnen de afbeelding. Elke bitmap bevat een vast aantal pixels. De afbeelding wordt verdeeld in groepen (arrays) van 8 naast of boven elkaar liggende pixels. Per array wordt bijgehouden welke pixel in welke kleur wordt weergegeven. Voor monochrome afbeeldingen volstaat 1 bit om een kleur weer te geven, maar voor kleuren en grijstinten vereist elk kleur meer dan één bit. De kleur wordt niet afzonderlijk voor elke pixel bijgehouden. In de array wordt aangegeven op welke pixel de kleur verandert en wat de nieuwe kleur is. Een bitmapafbeelding met grote oppervlakken dezelfde kleur zal dus minder bestandsomvang in beslag nemen dan een afbeelding met kleine oppervlakken. Een bitmapafbeelding is dus een samenstelling van arrays. De arrays worden voorafgegaan door computercode die eigen is aan het bestandsformaat zodat de computer weet dat de volgende bytes geen ASCII-karakters zijn. De

omzetting van bitpatronen in afbeeldingen kan soms enige tijd in beslag nemen. Bitmap- of rasterafbeeldingen hebben als algemene nadelen dat ze niet zonder vervorming schaalbaar zijn en resolutie-afhankelijk zijn. Bitmapafbeeldingen zijn het best geschikt om afbeeldingen met gradaties in kleuren en tinten te bevatten.

De bekendste bestandsformaten voor bitmap- of rasterafbeeldingen zijn TIFF, JPEG, GIF en PNG. Om een onderscheid te maken tussen de diverse rasterafbeeldingen zijn hun kleurdiepte en hun resolutie heel belangrijk. De kleurdiepte geeft weer hoeveel bits er worden gebruikt om de kleur van één pixel vast te leggen. De kleurdiepte bepaalt dus hoeveel verschillende kleuren een afbeelding maximaal kan bevatten. Hoe groter de kleurdiepte, des te meer kleuren de afbeeldingen kan bevatten (1-bit kleurdiepte: zwart/wit; 4-bit kleurdiepte: 16 kleuren; 8-bit kleurdiepte: 256 kleuren; 24-bit kleurdiepte: 16.777.216 kleuren). Elke bitmapafbeelding heeft een vaste resolutie. De resolutie is de dichtheid van de pixels die de afbeelding vormen. De resolutie bepaalt de scherpheid van de afbeelding (dpi: dots per inch (afdruk); ppi: pixels per inch (scherm); lpi: lines per inch (scanner)). Een resolutieverlaging is altijd mogelijk, voor een resolutieverhoging moet doorgaans het creatieproces worden overgedaan. De compressieloze bestandsomvang van een bitmapafbeelding wordt bepaald door de afmetingen en de kleurdiepte. Om de bestandsomvang in de hand te houden wordt doorgaans compressie gebruikt. Bitmap- of rasterafbeeldingen zijn dan ook binaire bestanden. Inkapseling van de nodige metadata in binaire bestanden is niet altijd evident.

De andere groep afbeeldingen zijn de vector- of object geörienteerde afbeeldingen. Bij deze groep wordt de afbeelding als een samenstelling van vormen opgeslagen. Door middel van wiskundige formules wordt van elke vorm de punten (x,y-coördinaten) bijgehouden. Vectorbestanden zijn gebaseerd op het verbinden van de lijnen tussen twee of meerdere punten. Zo ontstaan er vlakken en figuren waarvan de kleurwaarde wordt opgeslagen. In tegenstelling tot de rasterafbeeldingen ontstaat er geen vervorming bij schaling. Vectorafbeeldingen zijn evenmin resolutie-afhankelijk. Vectorafbeeldingen worden best gebruikt voor strak afgelijnde figuren. In vergelijking met rasterafbeeldingen nemen bestanden met vectorafbeeldingen meestal minder schijfruimte in beslag.

Het hierna volgende overzicht heeft voor het ogenblik enkel betrekking op tweedimensionele afbeeldingen. Het overzicht met de standaarden voor driedimensionele afbeeldingen of (technische) tekeningen is in voorbereiding.

C.3.1 Rasterafbeeldingen

C.3.1.1 Officiële standaarden

A) TIFF: Tagged Image File Format

Het TIFF-bestandsformaat (*.tif of *.tiff) werd ontwikkeld door Aldus Corporation en Microsoft Corporation, maar Aldus was eigenaar van de patenrechten. Na het samengaan van Aldus en Adobe Systems in 1994, gingen deze rechten over naar Adobe. In 1998 werd TIFF door ISO vastgelegd als officiële standaard: *ISO-12639: Graphic technology -- Prepress digital data exchange -- Tag image file format for image technology*. Er is ook een ISO-12234 voor digitale fotografie in de maak. De specificatie van het TIFF-formaat is vrij beschikbaar op de website van Adobe.

Er zijn verschillende versies van het TIFF-formaat. De eerste vastgelegde versie dateert van 1986 en kreeg het versienummer 3.0 (TIFF 1.0: draft 1; TIFF 2.0: draft 2). TIFF 4.0 werd in 1987 verspreid en bevat een aantal kleine wijzigingen. In 1988 werd al TIFF 5.0 vastgelegd. Deze versie ondersteunt paletkleuren en LZW-compressie. De laatste versie is TIFF 6.0 en dateert van 3 juni 1992. Met deze versie werden CMYK- en YC_BC_R-kleuren en de JPEG-compressie geïntroduceerd. TIFF 6.0 is tot op zekere hoogte compatibel met de vorige versies. De meeste computerprogramma's ontworpen voor versie 5.0 kunnen versie 6.0 inlezen, voor zover er geen gebruik is gemaakt van de specifieke uitbreidingen van TIFF 6.0. Momenteel is TIFF versie 7.0 in ontwikkeling, maar hierover werd nog geen informatie verspreid. Ondertussen is versie 6.0 nog steeds gangbaar, en kan het als een stabiel formaat worden beschouwd.

Een TIFF-bestand kan verschillende soorten stilstaande rasterafbeeldingen bevatten: bi-level, grijsschalen, RGB, YMCK, YC_BC_R en CIELab[30]. TIFF wordt veel gebruikt bij de opslag van ingescande afbeeldingen en foto's. TIFF-bestanden wordt ook veel gebruikt voor als opslagformaat voor tekstuele ingescande documenten. Afbeeldingen in TIFF kunnen in principe tot een kleurendiepte van 64 bits gaan, maar veel grafische applicaties ondersteunen maximaal 24 bits. Volgens de TIFF-specificatie is het mogelijk om meerdere afbeeldingen in één TIFF-bestand te bewaren, maar er zijn maar weinig applicaties die deze functionaliteit ondersteunen.

Een TIFF-bestand kan in principe maximaal 4 gigabytes groot zijn. Bij de opslag van afbeeldingen als TIFF-bestanden kan men ook compressie toepassen. Men heeft de keuze tussen geen compressie, CCITT-Groep 3 en 4, LZW, JPEG en Packbitscompressie[31]. CCITT-Groep 3 en 4 dient enkel voor bi-levelafbeeldingen. Compressieloze opslag en Packbitscompressie is altijd mogelijk bij de andere soorten afbeeldingen en behoren tot de baseline TIFF. LZW- en JPEG-compressie zijn uitbreidingen op de baseline TIFF, maar worden in de praktijk het meest gebruikt en zijn toepasbaar op alle modi.

TIFF heeft zijn naam te danken aan zijn samenstelling. De TIFF-bestanden bestaan uit velden (blokken) die geïdentificeerd worden door genummerde tags. Elk veld bevat gegevens van of over de afbeelding. Er zijn verplichte velden en optionele velden. De verplichte velden vormen de baseline TIFF. Alle TIFF-readers moeten in principe zowel de basis als optionele velden kunnen inlezen. Of bepaalde velden al dan niet voorkomen in het TIFF-bestand kan afhankelijk zijn van de toepassing waarmee TIFF-bestanden worden opgeslagen. Bij het ontwerpen van TIFF werd namelijk ook ruimte gelaten voor customialisering. De veldenstructuur van een TIFF-bestand maakt naast de basis- en optionele velden ook de opname van private velden mogelijk. Deze velden kunnen voor een specifiek gebruik dienen. Het gaat om de velden 32768 en hoger. Deze tags kunnen bij Adobe geregistreerd worden. Voorbeelden hiervan zijn: GeoTIFF (zie p. 49), TIFF voor PageMaker, Kodak TIFF, enz. Programma's kunnen de TIFF-blokken negeren die ze niet begrijpen of kunnen zich beperken tot het

---

[30] De verschillende soorten afbeeldingen:
- Bi-level of zwart-wit afbeeldingen: de pixel is zwart of wit (= modus bitmap in Photoshop)
- Grijsschalen of afbeeldingen met grijswaarden (4 of 8 bits): de pixel wordt weergegeven door een waarde tussen 0 (zwart) en 255 (wit). Er kunnen dus 256 verschillende grijstinten worden gebruikt (= modus grijswaarden in Photoshop).
- Paletkleur: één pixel wordt gevormd door één kleurstaal.
- RGB of Rood-Groen-Blauw afbeeldingen: per pixel wordt 24 bits (8 x 3) kleureninformatie opgeslagen. De pixel is samengesteld door een mengeling van de drie kleurstalen. RGB wordt door computermonitors toegepast. Afbeeldingen die bestemd zijn om op het scherm te worden bekeken (bijv. voor een website) worden bij voorkeur in RGB opgeslagen.
- CMYK of Cyaan-Magenta-Geel-Zwart afbeeldingen: elke pixel wordt samengesteld door deze vier kleuren. Afbeeldingen die worden afgedrukt worden best in CYMK-modus opgeslagen.

[31] Packbitscompressie is de run-length compressie die werd ontworpen door Apple.

inlezen van de blokken die ze nodig hebben. Dit biedt voor archivistische doeleinden de mogelijkheid om nieuwe velden aan een TIFF-bestand toe te voegen waarin bijvoorbeeld metadata worden opgenomen die niet tot de TIFF-basisvelden behoren (zie verder). Toch lijkt dit niet aangewezen te zijn. De vrijblijvende specificatie van het TIFF-formaat heeft ondertussen tot een wildgroei van TIFF-bestanden met verschillende interne structuren geleid die enkel met één bepaalde toepassing kunnen worden ingelezen[32]. Hierdoor neemt de platformonafhankelijkheid van het TIFF-formaat sterk af.. Een echt platformonafhankelijk TIFF-bestand is een bestand dat enkel uit de vereiste en optionele velden van de standaardspecificatie bevat.

Elk TIFF-bestand bestaat uit drie delen: de image file header (IFH), de image file directory (IFD) en de bitmap data. De eerste twee bytes van de image file header bepalen de byte orde. De waarde II (hex 49 49) geeft aan dat de *little-endian* volgorde werd toegepast en betekent dat de bytes van minst belangrijk naar meest belangrijk werd toegepast. Dit is de byte orde die door Intel-machines ("LSB") wordt toegepast. Wanneer daarentegen MM (hex 4D 4D) op de eerste twee posities van het TIFF-bestand staan, houdt dit aan dat de *big-endian* volgorde werd gevolgd. De bytes zijn hierin van belangrijkste naar minst belangrijkste gerangschikt. Dit is de byte orde van Mac-computers ("Motorola"). De volgende twee bytes van de IFH waren eigenlijk bedoeld als versienummer van het TIFF-bestand, maar geven eigenlijk enkel aan dat het om een TIFF-bestand gaat (de hexadecimale waarde 2A of decimaal 42). De laatste bytes van de IFH wijzen naar de beginpositie van de eerste IFD. De IFD bevat een beschrijving van de afbeelding en wijst naar de overeenstemmende bitmap data. In de IFD wordt onder andere de hoogte, breedte, de compressietechniek, de software en de datum en het tijdstip (jjjj:mm:dd uu:mm:ss) vastgelegd. De IFD points tenslotte naar de afbeeldingsdata. Eén TIFF-bestand kan meerdere image file directories en dus meerdere afbeeldingen bevatten. In het TIFF-bestand is er eveneens een blok voor de miniatuur van de afbeelding (thumbnail) voorzien.

Een aantal velden van het TIFF-bestand zijn tekstblokken waarin metadata over de afbeelding wordt opgenomen. Deze velden bevatten geen binaire data, maar gewone ASCII- of Unicodekarakters. Zo kan men beschrijvende gegevens aan het TIFF-bestand toevoegen zodat ze er een essentieel onderdeel van worden. Het gebruiken van deze velden voor metadata brengt de platformonafhankelijkheid van TIFF-bestanden niet in gevaar, aangezien het allemaal om basis- en optionele velden en niet om private velden gaat. De uitbaarheid van TIFF laat toe dat er bijkomende metadatavelden in het bestand worden voorzien, maar wegens de platformafhankelijkheid is dit niet aangewezen.

VELDNR.	BENAMING	WAARDE
TIFF: Baseline		
315	Artist	ASCII
259	Compression Schema	gebruikte compressie[33]

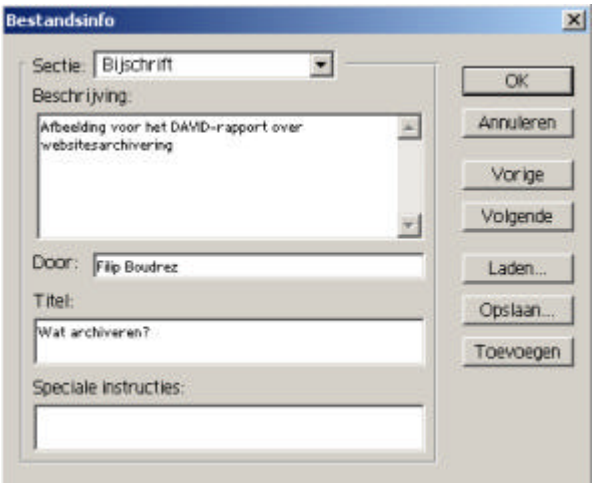
[32] In het kader van het Amerikaanse Defense Personnel Records Imaging System (DPRIS) werden de personeelsdossiers ingescand en als TIFF-bestanden bewaard. Hierbij werden verschillende TIFF-headers en uitbreidingen op het TIFF-formaat gebruikt, waardoor de uitwisseling van deze bestanden heel moeizaam verloopt (S. MACTAVISH, *DoD-NARA Scanned Images Standards Conference*, in: RLG Diginews, april 15, 1999, vol. 3, nr. 2.)

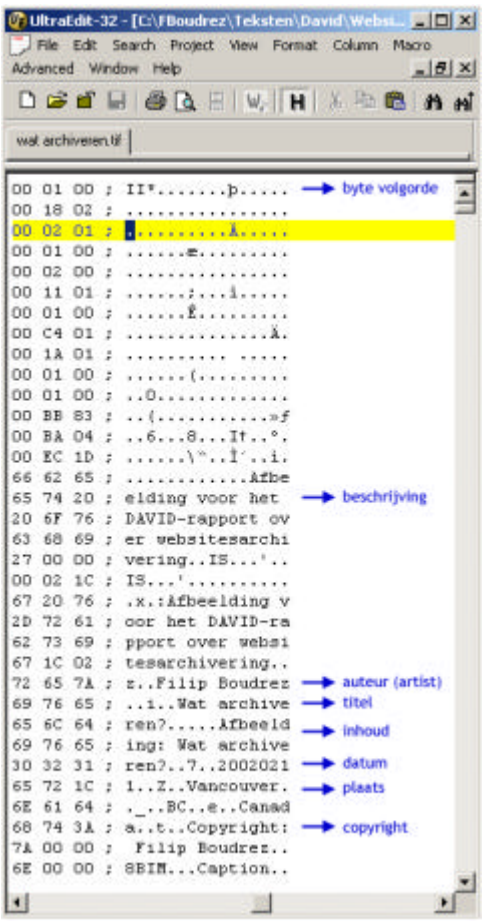
[33] Het veldnummer 259 wordt gevolgd door een cijfer. Elk cijfer staat voor een bepaalde compressie

1:	geen compressie	5:	LZW
2:	CCITT-groep 1 D	6:	JPEG
3:	CCITT-groep 3	32773:	Packbits compressie

33432	Copyright	ASCII
306	DateTime	ASCII
316	HostComputer	computer en OS die werd gebruikt bij de creatie van het bestand
270	ImageDescription	ASCII-waarde
271	Make	fabrikant van de scanner, video, apparatuur die werd gebruikt voor het maken van de afbeelding
272	Model	model van de scanner die werd gebruikt
305		software naam en versienummer van de software die werd gebruikt voor het maken van de afbeelding
TIFF:	Extensions	
269	DocumentName	ASCII
285	PageName	ASCII
297	PageNumber	ASCII

Niet elke applicatie echter laat het toevoegen of bekijken van deze metadata toe. Photoshop laat de invulling van een aantal velden toe. In gewone grafische programma's zoals ACDSee of Microsoft PhotoEditor kan enkel de beschrijving worden opgevraagd. Dit vormt echter geen enkel probleem want in een gewone teksteditor kunnen deze velden altijd bekeken worden aangezien ze uit gewone ASCII- of Unicodekarakters bestaan.





Afbeelding 3 en 4: Via de opdracht 'Bestandsinfo' in het menu 'Bestand' in PHOTOSHOP is het mogelijk om metadata aan het TIFF-bestand toe te voegen. Deze metadata worden opgeslagen als ASCII-karakters en zijn dus raadpleegbaar met een gewone teksteditor.

4: CCITT-groep 4

In de laatste TIFF-versie zijn diverse uitbreidingen voorzien. Deze zijn onder meer: alfakanalen, CCITT-4 en CCITT-6 compressie voor bi-level afbeeldingen, YcbCr-afbeeldingen, JPEG-compressie en CIELab. Deze TIFF-uitbreidingen worden echter niet door alle grafische programma's ondersteund. Zo zijn er een aantal programma's die enkel met de baseline TIFF compatibel zijn.

TIFF is initieel ontworpen voor desktop publishing en met de bedoeling uitwisseling van rasterafbeeldingen mogelijk te maken. TIFF-bestanden zijn uitwisselbaar tussen MSDOS, Unix, Mac en IBM-machines. De byte orde waarin TIFF-bestanden worden opgeslagen, doet er eigenlijk niet echt toe. Volgens de TIFF-specificatie moeten TIFF-readers beide byte ordes kunnen inlezen. TIFF-bestanden zijn platform- en bestandssysteemonafhankelijk en bijgevolg uitwisselbaar. TIFF is uitgegroeid tot één van de basisbestandsformaten waarin rasterafbeeldingen worden opgeslagen. Bijna alle tekstverwerkings-, teken- of paginabewerkingsprogramma's kunnen TIFF-bestanden openen. Er is in het publiek domein voor elk platform en besturingssysteem een heel gamma applicaties beschikbaar dat TIFF-afbeeldingen kunnen lezen. TIFF-bestanden dienen niet voor vectorafbeeldingen.

De masters van gedigitaliseerde archiefdocumenten (foto's, briefwisseling, plannen, enz.) worden doorgaans in TIFF opgeslagen. Op basis een van TIFF-bestanden kan men van de afbeelding nog een versie in een ander bestandsformaat bewaren. Bij het bewaren van masters wordt best ook geen compressie toegepast. Bij masters wordt er best ook een hoge resolutie gebruikt zodat ze bij eventueel later gebruik niet opnieuw moeten worden ingescand. Een veel gebruikte resolutiewaarde is 300 dpi.

TIFF wordt niet ondersteund door de courante webbrowsers. Hiervoor is een plug-in nodig.

Photoshop slaat de lagen in een TIFF-afbeelding afzonderlijk op. Wanneer de afbeelding vervolgens in een andere applicatie wordt geopend, dan worden de lagen samengevoegd. In de andere applicaties kunnen de verschillende lagen bijgevolg niet afzonderlijk worden bewerkt.

Referentie: TIFF: Revision 6.0. Final, juni 1992.

B) JPEG: Joint Photographic Experts Group

Het JPEG-bestandsformaat is genoemd naar de groep experten (Joint Photographic Experts Group) die door nationale standaardiseringsorganisaties en belangrijke bedrijven werd samengesteld met als doel een gemeenschappelijke compressiestandaard voor afbeeldingen in grijsschalen en kleuren te produceren. De JPEG-groep werkt onder de vleugels van ITU, ISO en IEC. Het JPEG-initiatief werd in de jaren 1980 opgestart en wordt nog steeds verder gezet. De benaming JPEG verwijst eigenlijk in de eerste plaats naar de JPEG-compressiemethode. Deze techniek werd in allerhande applicaties verwerkt (bijv. bij opslag van afbeeldingen in PDF-documenten, compressie van een TIFF-bestand). Met de benaming JPEG wordt ten tweede ook het bestandsformaat aangeduid. Dit bestandsformaat is gebaseerd op de JPEG-compressie met daar rond een file wrapper. Inmiddels werden JPEG-LS en JPEG-2000 in 2000 ook door ISO als officiële standaard gepubliceerd. De JPEG-groep is nauw verwant met JBIG. Deze laatste groep werkt op bi-level en gereduceerde grijstintenafbeeldingen.

B. 1 JPEG

Het JPEG-bestandsformaat werd in augustus 1990 vastgelegd als onderdeel van de ISO-10918 standaard (*ISO-10918-4(1999): Information Technology. Digital Compression and coding of continuous-tone still images: Registration of JPEG profiles, SPIFF profiles, SPIFF tags, SPIFF colour spaces, APPn markers, SPIFF compression types and Registration Authorities (REGAUT)*).. Dit is een multipart standaard voor compressie voor stilstaande afbeeldingen.

Deze JPEG-versie wordt door technici ook JPEG-DCT genoemd. DCT (Discrete Cosine Transform) verwijst naar de gebruikte compressiemethode gebaseerd op de (co-)sinusgolven. Binnen JPEG worden meerdere modes van elkaar onderscheiden. Er zijn twee basismodes: lossy (baseline) en lossless. Beide modes gebruiken de Huffman-coding, maar het gebruikte compressie-algoritme is helemaal anders. Binnen JPEG zijn er ook een aantal varianten. Deze varianten zijn op de twee basismodes gebaseerd. Hierarchische en progressieve JPEG (beide lossy) zijn de bekendste varianten.

De lossless JPEG-basismode (L-JPEG) wordt nauwelijks of niet toegepast en is enkel in gespecialiseerde computerprogrammatuur ingebouwd. De lossy JPEG-basismode is daarentegen zeer populair en is in bijna alle grafische computerprogramma's geïmplementeerd. JPEG wordt bijvoorbeeld heel veel gebruikt bij scanning en digitale fotografie. Wanneer men het in de omgangstaal over JPEG-bestanden heeft, dan bedoelt men meestal deze mode. In de baseline mode wordt elke afbeelding in blokken van 8 x 8 pixels opgedeeld. Elke blok wordt getransformeerd door DCT. De mate waarin men informatie bij de compressie verliest, is aanpasbaar door de parameters in te stellen. De JPEG-compressie en het bijgaand verlies van informatie is op de waarnemingsmogelijkheden van het menselijk oog gebaseerd. Elementen die de mens toch niet kan waarnemen, worden als eerste uit de afbeeldingen selectief weggefilterd. Hoe groter de compressieratio, hoe meer informatie verloren gaat. Heel typisch hiervoor is het verdwijnen van fijne details (bijv. tekstuele informatie in een afbeelding). Deze JPEG-versie is in vergelijking met wavelet of fractal compressie niet zo efficiënt. Een afbeelding met dezelfde kwaliteit is met lossy JPEG-compressie twee of drie maal groter dan een bestand gecomprimeerd met wavelet of fractale compressie.

Voor JPEG-bestanden wordt een 24-bits kleurdiepte gebruikt. JPEG dient dan ook in de eerste plaats voor de opslag van kleuren (truecolor: RGB, CMYK, (YC_BC_R) en zwart-wit (grijsschaal) foto's. JPEG ondersteunt geen transparantie.

JPEG wordt het meest in netwerk- of webomgevingen gebruikt. Hierbij wordt meestal een progressief JPEG-bestand opgeslagen. Dit betekent dat de JPEG-afbeelding in een lage resolutie integraal wordt getoond, terwijl ondertussen de rest van de afbeelding wordt gedownload. De courante grafische programma's bieden de mogelijkheid om progressieve JPEG-bestanden te maken. JPEG wordt door de meeste webbrowsers ondersteund.

Voor archiveringsdoeleinden worden JPEG-bestanden vooral in combinatie met TIFF-bestanden gebruikt. Bij het digitaliseren van afbeeldingen worden de hoge resolutieversies in TIFF bewaard. De versies die (on line) ter beschikking worden gesteld, worden als lage resolutie JPEG-bestanden verspreid.

Een bekend nadeel van JPEG is het optreden van vervormingen zoals golven en geblokte strepen bij veelvuldige bewerkingen op basis van hetzelfde JPEG-bestand. Beter is om de bewerking op (een duplicaat van) de master uit te voeren en de bewerking vervolgens als een JPEG-bestand te bewaren.

In een JPEG-bestand worden ook een aantal metadata opgeslagen: o.a. tijdstip en datum van creatie, capturing device of computerprogramma, bestandsomvang, auteursrecht.

B. 2 JPEG-LS

JPEG-LS is de recentste ISO/ITU-T standaard voor lossless coding van stilstaande afbeeldingen. (*ISO-144495 (2000): Information technology -- Lossless and near-lossless compression of continuous-tone still images: Baseline*). Deze JPEG-versie biedt naast lossless compressie ook een "near-lossless" compressiemethode aan.

B.3 JPEG-2000

Op het einde van 2000 werd ook de lang aangekondigde JPEG-2000 (*.jp2, *.j2k,) als officiële standaard gefinaliseerd (*ISO/IEC 15444-1(2001): Information technology -- JPEG 2000 image coding system -- Part 1: Core coding system*). JPEG-2000 is bedoeld als een uitbreiding en verfijning van de bestaande compressiestandaarden voor de opslag van stilstaande beelden. Deel 1 is al gepubliceerd en is volledig vrij van patentrechten. Dit deel heeft betrekking op de kern van de standaard. Deel 2 is momenteel nog in ontwikkeling en zal uitbreidingen voor specifieke toepassingen (behandeling van tekst, animatie-effecten, metadata) bevatten. Deel 2 zal meer dan waarschijnlijk door auteursrechten worden beschermd.

Eén van de belangrijkste vernieuwingen van JPEG-2000 is het gebruik van een waveletcompressiemethode. Deze methode wordt Discrete Wavelet Transform (DWT) genoemd. Bij zijn publicatie werd gesteld dat hiermee een compressieratio van 200:1 wordt bereikt, maar dit zal in de praktijk zelden haalbaar zijn. De compressieratio van JPEG-2000 ligt wel beduidend hoger dan bij de eerste JPEG-versie. JPEG-2000 omvat net zoals JPEG een lossless en een lossy compressiemethode. JPEG-2000 werkt op basis van blokken van 64 x 64 pixels. JPEG-2000 kan in principe eveneens met lossless compressie worden toegepast.

JPEG-2000 biedt naast de wavelet compressiemethode ook een aantal andere nieuwe functionaliteiten aan:

- kleurenbeheer: paletkleuren, (s)RGB, CYMK, YC_BC_R, ICC
- mogelijkheid om 'regions of interest' aan te duiden: bepaalde delen van een afbeelding kunnen lossless worden opgeslagen terwijl voor de rest van de afbeelding lossy compression wordt toegepast
- gebruik van alpha channels (transparantie)
- toevoegen van metadata die in het JPEG-bestand worden ingekapseld
- grotere kleurdiepte
- hoog fout herstellingsvermogen
- random access
- schaalbaarheid

Het gevolg van deze uitbreidingen is natuurlijk dat de JPEG-2000 complexer is dan bijvoorbeeld de JPEG-oerversie. Net zoals alle andere JPEG-versies bereikt JPEG-2000 een hogere compressieperformantie bij lossy compressie dan bij lossless compressie.

Het is momenteel wachten op de implementatie van deze nieuwe officiële standaard in computerprogramma's. Voor het bekijken van JPEG-2000 afbeeldingen in een webbrowser is een plug-in vereist.

Referentie: http://www.jpeg.org

C.3.1.2 Defacto standaarden

A) BMP

BMP is het standaardformaat voor afbeeldingen dat standaard door de besturingssystemen DOS en Windows wordt gebruikt. Een BMP-afbeelding bestaat uit vier delen: een header, een informatieheader, een kleurentabel en de data van de eigenlijke afbeelding. BMP-afbeeldingen kunnen verschillende kleurdieptes hebben: 1-bit (zwart-wit), 4-bits (16 kleuren), 8-bits (256 kleuren) en 24-bits (16,7 miljoen kleuren). Een BMP-bestand kan zowel monochrome, geïndexeerde kleuren, grijswaarden als RGB-kleuren bevatten. BMP-afbeeldingen bevatten geen alfakanalen. Versie 1 (*Device Dependant Bitmap*) gebruikt geen compressie en is gebaseerd op een vast kleurenpalet. Versie 2 (*Device Independant Bitmap*; vanaf Windows 3.0) gebruikt RLE-compressie en het kleurenpalet is manipuleerbaar. De 1-bit en 4-bits BMP-afbeeldingen gebruiken RLE-4 compressie, terwijl de 8-bits en 24-bits afbeeldingen met behulp van RLE-8 compressie worden opgeslagen. Ondanks de compressie hebben BMP-afbeeldingen over het algemeen een vrij grote bestandsomvang.

Referentie: /

B) GIF: Graphics Interchange Format

GIF is één van de oudste bestandsformaten voor afbeeldingen. De GIF-specificatie gaat terug tot 1987. GIF was het enige afbeeldingsformaat dat door de eerste generatie grafische webbrowsers werd ondersteund.

Er zijn twee wijdverspreide versies van GIF. De versie GIF 87a bevat één stilstaand beeld. De versie GIF 89a kan ook een sequentie van opeenvolgende statische beelden (frames) bevatten waardoor beweging wordt gesimuleerd (animated gifs). Het versienummer staat als ASCII-karakters op het begin van het bestand.

Een GIF-afbeelding kan voor elke pixel slechts 8-bits aan kleurinformatie bevatten. Hierdoor is het aantal kleuren in een GIF-afbeelding beperkt tot maximaal 256 paletkleuren. Aan de andere kant laat GIF ook toe dat de gebruiker het aantal kleuren in een GIF-afbeelding beperkt. Men doet dit doorgaans om een kleinere bestandsomvang te bereiken.

GIF is uitermate geschikt om lage resolutie afbeeldingen in paletkleuren met effen vlakken en details te bevatten (bijv. monochrome afbeeldingen, tekeningen en cartoons). GIF is helemaal niet ontworpen om foto's in kleur of in grijswaarden in op te slagen. In een GIF-afbeelding worden geen

alfakanalen ondersteund. Eén niveau transparantie (transparant of volledig bedekt) in afbeeldingen blijft daarentegen wel behouden. GIF-bestanden zijn onafhankelijk van de hardwareconfiguratie waarop ze werden gecreëerd.

Als men het GIF-formaat gebruikt voor afbeeldingen waarvoor het eigenlijk is ontworpen, treedt er bijna geen informatieverlies op. Het informatieverlies is geen gevolg van de lossless LZW-compressie die altijd wordt toegepast. De LZW-compressie is in de eerste plaats gericht op grote effen vlakken in een afbeelding. Bij een eigenlijk gebruik van GIF-bestanden is de compressieratio heel doeltreffend. Hierdoor hebben GIF-bestanden bijna altijd een relatief kleine bestandsomvang en kunnen ze relatief gemakkelijk via netwerken getransporteerd worden. Het optreden van informatieverlies is meestal een gevolg van de beperking op het aantal kleuren. GIF is beperkt tot een kleurdiepte van 8 bits. Het is bijgevolg evident dat er bij het bewaren van een 24 bits RGB-afbeelding als GIF-bestand veel kleuren of schakeringen verloren gaan.

Bij het bewaren van een afbeelding als GIF-bestand kan men in de meeste grafische programma's de dithering bepalen en de rijvolgorde specifiëren. Dithering is een manier waarop ontbrekende kleuren in de kleurentabel worden gesimuleerd. De rijvolgorde bepaalt hoe een afbeelding in een webbrowser wordt weergegeven. Er is doorgaans de keuze tussen 'normaal' en 'geïnterlinieerd' (interlacing). Bij 'normaal' wordt eerst de volledig afbeelding gedownload en dan pas op het scherm getoond. Bij 'geïnterlinieerd' wordt er eerst een volledige afbeelding in lage resolutie getoond, terwijl de resolutie verfijnt tijdens het verder downloaden (vergelijkbaar met progressieve JPEG). Deze laatste rijvolgorde heeft wel een grotere bestandsomvang als gevolg.

De GIF-specificatie voorziet een vrij tekstveld waarin de gebruiker in principe welke informatie over de afbeelding zou kunnen opnemen.

Een aantal jaren geleden was er heel wat te doen rond het GIF-bestandsformaat. GIF is uitgewerkt door CompuServe. In de overtuiging dat het LZW-algoritme tot het publiek domein behoorde, heeft CompuServe de compressie van GIF-bestanden hierop gebaseerd. De problemen rezen toen Unisys zijn eigendomsrechten op de LZW-compressiemethode liet gelden. Unisys is houder van de patentrechten op het algoritme en de licensies van grafische programma's die GIF-bestanden creëren dekken de LZW-licensie niet. Voor het bewaren van archiefdocumenten als GIF-bestanden moet dus in principe een licentie bij Unisys worden aangeschaft. Met andere woorden, de licentierechten voor Photoshop hebben enkel betrekking op het gebruik van het programma en niet op de creatie van GIF-bestanden. Uit onvrede met deze gang van zaken werd PNG (zie hieronder) als alternatief gecreëerd. Niettegenstaande dit gegeven is GIF tot op de dag van vandaag een heel populair bestandsformaat in Internet- en netwerktoepassingen.

Referentie: *G I F: Graphics Interchange Format. A standard defining a mechanism for the storage and transmission of raster-based graphics information*, 1987 (CompuServe Incorporated, 1987); *Graphics Interchange Format* (versie 89a).

C) PNG: Portable Network Graphics

PNG is ontworpen door het *World Wide Web Consortium* om een antwoord te bieden op de licensieproblemen rond GIF en zijn LZW-compressie. PNG gebruikt bijgevolg een andere

compressiemethode dan GIF (LZ77 en Guffman). PNG heeft een aantal gelijkenissen met GIF, maar is in veel opzichten een verfijning en uitbreiding. PNG-bestanden zijn hardware-onafhankelijk. PNG werd op 1 oktober 1996 door het WWW als Recommendation gepubliceerd.

Net zoals GIF is PNG een bestandsformaat voor stilstaande afbeeldingen. Bij PNG is er de keuze tussen het gebruiken van een 8-bits of 24-bits kleurdiepte. PNG-8 dient voor dezelfde soort afbeeldingen met paletkleuren die in GIF kunnen worden opgeslagen. Het compressieschema van PNG-8 is verfijnder dan dat van GIF. Dezelfde afbeelding opgeslagen in PNG-8 kan 10 tot 30 % kleiner zijn. De enige uitzondering hierop zijn afbeeldingen met weinig kleuren en eenvoudige patronen. De PNG-compressie is een lossless compressiemethode. Net zoals bij GIF kan men bij PNG de dithering en het maximale aantal kleuren bepalen.

PNG-24 bevat meer kleuren dan GIF en is geschikt om zowel (s)RGB/ICC-afbeeldingen als afbeeldingen met grijswaarden te bewaren. Inzake transparantie biedt PNG-24 voor elke pixel 256 niveau's. PNG-24 gebruikt dezelfde compressiemethode als PNG-8. Deze compressiemthode is echter niet zo geschikt voor afbeeldingen met ware kleuren, hoge kleuren of grijstinten. Dezelfde afbeelding opgeslagen in JPEG heeft doorgaans een kleinere bestandsomvang. PNG bereikt de hoogste compressieratio op afbeeldingen met grijswaarden en scoort hier beduidend beter dan de recentste formaten zoals JPEG-LS, JPEG-2000 en MPEG-4 VTC.

PNG heeft nog als kenmerken de ondersteuning van alfakanalen, gamma correctie en twee dimensionele interlacing. Eén pixel in een PNG-bestand kan een variatie in transparantie of dekking van 256 niveau's hebben (alfakanalen). Gammacorrectie verbetert de verschillende kleureninterpretaties van computers (bijv. de verschillen in lichtsterkte). Ten slotte wordt een PNG-bestand sneller op het scherm weergegeven dan een GIF-bestand.

PNG-afbeeldingen kunnen geen animatie bevatten zoals GIF89a. De W3C-tegenhanger van animated gifs is *Multiple-image Network Graphics*.

PNG heeft het statuut van open specificatie binnen de groep van defacto standaarden. PNG is momenteel ook het onderwerp van de standaardisatieprocedure binnen ISO/IEC JTC1/SC24 en zal wellicht een officiële standaard worden: ISO/IEC-15948.

Tot op heden kent PNG nog maar een kleine toepassing. PNG wordt vooral binnen Internettoepassingen gebruikt. PNG is bijvoorbeeld het native bestandsformaat van Macromedia's Fireworks[34]. Binnen andere soorten grafische applicaties is PNG nog nauwelijks of niet ingeburgerd. PNG wordt ook nog niet door alle webbrowsers ondersteund, maar wel al door Internet Explorer en Netscape. Voorbeelden van het gebruik van PNG bij digitale archivering zijn ons niet bekend. PNG heeft immers een aantal nadelen bij gebruik voor archiveringsdoeleinden. Bij PNG wordt altijd compressie toegepast en de PNG-specificatie laat nagenoeg geen ruimte open bij implementatie. PNG is enkel geschikt voor afbeeldingen in RGB en grijswaarden en niet voor CYMK of apparaatonafhankelijke kleurschema's.

Referentie: http://www.w3.org/tr/rec-png.html; http://www.w3.org/tr/png.html; http://www.libpng.org/pub/png/

[34] De PNG-bestanden die met Fireworks worden gemaakt zijn echter niet 100 % conform de PNG-specificatie samengesteld. De Fireworks PNG-bestanden bevatten een aantal bijkomende elementen die niet in het formele PNG-formaat zijn voorzien. Omzettingen van Fireworks naar andere applicaties of andere formaten kunnen hierdoor tot moeilijkheden in informatieverlies leiden.

D) Encapsulated Postcript

Een encapsulated postscriptbestand (*.eps) is eigenlijk ontworpen om de uitwisseling van postscriptbestanden tussen verschillende computerplatformen mogelijk te maken. Hierdoor wordt EPS beschouwd als een apparaatonafhankelijk bestandsformaat waardoor het geschikt is voor de uitwisseling van bestanden. Voor het openen van EPS-bestanden is wel een postscriptinterpreter nodig. De huidige versie is EPS 3.0 die in mei 1992 werd bekend gemaakt.

Een EPS-bestand kan zowel tekst, grafieken als afbeeldingen bevatten. In de meeste gevallen wordt in een EPS-bestand een afbeelding of één blad beschreven die in een ander document wordt opgenomen. EPS wordt bijgevolg als een grafisch bestandsformaat beschouwd. Het EPS-bestand bevat doorgaans de afbeelding die aan een bestaande postscriptpagina wordt toegevoegd. EPS kan zowel een vector- als bitmafbeelding bevatten (Lab, CMYK, RGB, geïndexeerde kleur, duotoon, grijswaarden en bitmap).

In een EPS-bestand kan hoogstens de verschijningsvorm van één pagina worden beschreven. Een EPS-bestand moet voldoen aan de Adobe Document Structuring Conventions (DSC). Het bestand moet minstens een header en een boundingbox bevatten. De boundingbox beschrijft de afmetingen en plaats van de afbeelding.

Een EPS-bestand bevat doorgaans een preview of een thumbnail. De gebruiker krijgt deze preview op het scherm te zien zodat kleine transformaties en positioneringen mogelijk zijn. Deze preview is doorgaans wel machine-afhankelijk. Elk besturingssysteem heeft zijn voorkeur voor een bepaald bestandsformaat. Voor Apple MacIntosh is dit PICT, voor Windows TIFF. Er is echter ook de mogelijkheid om de preview als een puur ASCII-bestand op te nemen. Een dergelijk EPS-bestand wordt een EPSI-genoemd. De printer drukt evenwel het EPS-bestand af, en niet de ASCII, TIFF of PICT-beeldschermversie.

Een EPS-bestand kan zowel ASCII-karakters als binaire data bevatten. Vanwege de platformonafhankelijkheid wordt het gebruik van binaire data echter afgeraden en beperkt men zich volgens de ESP-specificatie best tot 7 bit ASCII.

EPS heeft dezelfde voordelen als een postscript-bestand. Het verschil met een gewoon postscriptbestand is de toevoeging van commentaren. Een EPS-bestand kan gecreëerd worden door met behulp van een teksteditor of tekstverwerker de nodige code aan het postscriptbestand toe te voegen. Een andere mogelijkheid is het gebruik van grafische programma's die afbeeldingen als een EPS-bestand kunnen opslagen.

EPS wordt ondersteund door de meeste grafische programma's, tekenprogramma's en pagina-opmaakprogramma's. EPS kan bijvoorbeeld gebruikt worden als archiveringsformaat voor een afbeelding die in CorelDRAW werd gemaakt omdat het alle effecten overneemt.

Referentie: http://partners.adobe.com/asn/developer/pdfs/tn/5002.EPSF_Spec.pdf

E) GeoTIFF

Het GeoTIFF formaat is gebaseerd op het TIFF-rasterformaat en wordt gebruikt voor afbeeldingen binnen cartografische toepassingen. GeoTIFF werd ontwikkeld met de bedoeling een niet-producent gebonden oplossing te bieden voor de uitwisseling van cartografische afbeeldingen. Producenten zoals Intergraph, ESRI en Island Graphics bieden hier wel een oplossing voor, maar deze zijn eigendomsgebonden en blijven beperkt tot de noden van de eigen software. GeoTIFF behoort tot het publiek domein en is vrij van licentie-of patentrechten.

Voor de uitwisseling van cartografische afbeeldingen werd besloten zich op het TIFF-formaat versie 6.0 te baseren. TIFF is uitwisselbaar en platformonafhankelijk, kan als een heel stabiel bestandsformaat worden beschouwd en behoort tot het publiek domein. TIFF is één van de weinige rasterformaten dat bruikbaar is voor alle typen afbeeldingen die in de geografie wordt gebruikt. Bovendien biedt de TIFF-specificatie de mogelijkheid om metadata samen met de eigenlijke afbeeldingsdata in één en hetzelfde bestand op te nemen. GeoTIFF maakt gebruik van de zogenaamde "private" of "gereserveerde" TIFF-tags voor de opslag van georeferentiële en andere cartografische metadata van de afbeelding. Hierdoor worden 6 tags gebruikt. Deze metadata zorgen binnen een GeoTIFF compatibele toepassing ondermeer voor automatische lokalisatie en schaling

GeoTIFF-bestanden zijn net zoals TIFF-afbeeldingen uitwisselbaar. GeoTIFF biedt dezelfde voordelen als TIFF. Ook afbeeldingsverwerkingstoepassingen die niet met GeoTIFF compatibel zijn, maar wel met TIFF kunnen de afbeelding/kaart als een gewone TIFF-afbeelding openen. Deze programma's hebben wel geen toegang tot de geodata. De meeste GIS-toepassingen ondersteunen GeoTIFF.

De GeoTIFF specificatie behoort tot het publiek domein. Het is eveneens ook de bedoeling dat het publiek betrokken is bij het samenstellen, herzien of uitbreiden van het formaat. De verdere ontwikkeling van GeoTIFF brandt de laatste tijd echter op een laag pitje.

Referentie: http://remotesensing.org/geotiff/geotiff.html

F) FlashPix

Het FlashPix-bestandsformaat is in ontstaan uit de samenwerking tussen de vier computergiganten Hewlett-Packard, Kodak, Live Picture en Microsoft Corporation. In juni 1996 werd het nieuwe bestandsformaat officieel voorgesteld. Het beheer van FlashPix is nu in handen van de Digital Imaging Group (DIG). DIG is een onafhankelijk consortium dat ondermeer het Internet Imaging Protocol reguleert. FlashPix is een open industrie standaard.

Het bestandsformaat heeft een aantal eigenschappen die heel attractief lijken vanuit archivistisch perspectief: één computerbestand bevat dezelfde afbeelding op verschillende resoluties, wijzigingen worden samen met het origineel in hetzelfde bestand bewaard en de gebruiker kan tekstuele metadata aan het bestand toevoegen. FlashPix-bestanden hebben als extensie *.fpx. FlashPix-afbeeldingen hebben een bitdiepte van 8 bits (grijsschalen) of 24 bits (volle kleuren). Het kleurenbeheer is gebaseerd op PhotoYCC, NIF RGB en ICC.

Een FlashPix bestand is opgebouwd als een piramide. De afbeelding met de grootste resolutie vormt de basis van de piramide. Op de hogere lagen van de piramide staat dezelfde afbeelding, telkens in een lagere resolutie, die een kwart is van de vorige (bijv. 1000 x 800 pixels, 500 x 400 pixels, 250 x

200 pixels, enz.). Elke resolutie vormt een afzonderlijke array die binnen het FlashPix-bestand aan elkaar gekoppeld zijn op basis van OLE. FlashPix compatibele applicaties kunnen de afbeelding in één bepaalde laag inlezen. Deze mogelijkheid biedt het voordeel dat men voor een digitaal afbeeldingenarchief dat on line raadpleegbaar is, men de moederkopieën op een hoge resolutie en de afgeleide Internetversies niet meer als afzonderlijke bestanden moet beheren. Hetzelfde FlashPix-bestand bevat zowel de moederkopie als dezelfde afbeelding op een lagere resolutie. Voor terbeschikkingstelling via Internet kan de afbeelding op een lagere resolutie worden verspreid of kan men bijvoorbeeld op basis van die resolutie een JPEG-versie on the fly genereren. Aangezien FlashPix een multi-resolutie formaat is, kan er ook probleemloos op een detail van de afbeelding worden ingezoemd zonder dat de afbeelding kwaliteit verliest.

Elke laag is onderverdeeld in tegels van 64 x 64 pixels. Elke tegel kan afzonderlijk gecomprimeerd of ongecomprimeerd zijn. Voor compressie wordt JPEG-compressie of single color compressie toegepast. De FlashPix compressie is dus lossy. De bronbestanden van de FlashPixafbeeldingen kunnen zowel GIF, PNG, JPEG als TIFF zijn. FlashPix dient vooral voor afbeeldingen met grijswaarden of ware kleuren (24-bits kleurdiepte). Doordat FlashPix-bestanden dezelfde afbeelding op verschillende resoluties bevat, is ongecomprimeerd de bestandsomvang ongeveer 1/3 groter in vergelijking met een TIFF-afbeelding. FlashPix-compatibele applicaties kunnen elke tegel afzonderlijk inlezen. Bij het bekijken, aanpassen of afdrukken van een deel van de afbeelding worden enkel de nodige tegels ingelezen en niet de volledige afbeelding, wat de snelheid aanzienlijk vergroot en het computergeheugen minder belast (volgens Kodak tot minder dan 20 % van het RAM-geheugen).

Een FlashPix-bestand kan ook verschillende versies van dezelfde afbeelding bevatten. Als men bij andere bestandsformaten origineel en wijziging wil archiveren, moeten er minstens twee bestanden worden bewaard. In een FlashPix-bestand daarentegen wordt enerzijds de originele afbeelding opgeslagen en anderzijds de wijzigingen in de vorm van een script of batchoperatie. Als een gewijzigde versie wordt opgevraagd, wordt eerst de afbeelding op de overeenstemmende resolutie geopend en vervolgens worden de wijzigingen één na één uitgevoerd. De gewijzigde afbeelding wordt dus niet als een afzonderlijk bestand opgeslagen. De wijzigingen worden als het ware als verschillende scripts opgeslagen die vervolgens na opening in real time worden uitgevoerd. Hierdoor is er beduidend minder opslagcapaciteit nodig en is het gemakkelijker om de afbeeldingsbestanden te beheren.

Een FlashPix-bestand kan eveneens de nodige metadata bevatten. Door middel van toepassing van het OLE-principe kunnen textuele metadatagegevens worden toegevoegd aan het bestand. In deze gestructureerde velden kunnen ASCII-karakters worden opgenomen. Een aantal vaste metadatavelden zijn: auteursrecht, inhoud, capture device information, camera settings, device characterizations, beschrijving van de film, originele scan, scan apparaat. Aangezien FlashPix aanpasbaar is volgens de behoeften van de gebruikers, kunnen de metadatavelden uitgebreid worden. Op die manier worden de metadata in het afbeeldingsbestand zelf ingekapseld, en is men minder afhankelijk van externe databanken. De vereiste hiervoor is dat de computertoepassing toegang heeft tot de structuur van het bestandsformaat.

Ondanks al deze interessante voordelen, kent FlashPix nog maar een kleine verspreiding. Dit heeft zijn redenen. De ondersteuning van het bestandsformaat is nog maar heel beperkt. De meeste grafische computerprogramma's en databanken voor afbeeldingen ondersteunen dit formaat nog niet. De webbrowsers Internet Explorer en Netscape hebben een bijzondere plug-in nodig om de afbeeldingen te openen. Nochtans is er de toezegging van de grote computerbedrijven om het bestandsformaat in hun produkten te implementeren. Het W3C heeft Flashpix nog niet als WWW-standaard aanvaard (wel GIF, PNG en JPEG). Voor de creatie van FlashPix-bestanden is gespecialiseerde software nodig.

Binnen de musea, bibliotheken en archieven is er bijgevolg nog maar weinig ervaring met FlashPix. Ondanks deze minpunten blijft het afwachten hoe de FlashPix-ondersteuning in de toekomst evolueert.

Referentie: http://www.kodak.com/US/en/digital/flashPix/index.shtml

G) ImagePAC

Het ImagePac-formaat is het bestandsformaat waarin afbeeldingen op een Kodak Photo-CD worden opgeslagen. Kodak Photo-CD is een technologie die in 1990 werd gelanceerd met de bedoeling om afbeeldingen en in het bijzonder foto's digitaal op CD te plaatsen. De digitale foto's konden bekeken worden op computer met een CD-ROM XA drive of op televisie met behulp van een CD-I toestel of Photo-CDspeler.

De technologie was erop gericht om de inhoud van een traditionele 35 mm. fotofilm te digitaliseren en op CD te plaatsen. Een Kodak Photo-CD bevat doorgaans 100 afbeeldingen. Daarnaast bevat de CD een bestand met een overzicht van alle thumbnails en wat software.

Het ImagePac-formaat kan afbeeldingen in zes verschillende resoluties bewaren: 128 x 198 pixels, 256 x 384 pixels, 512 x 768 pixels, 1024 x 1536 pixels, 2048 x 3072 pixels en 6144 x 4096 pixels[35]. Dit heeft voor gevolg dat de originele afmetingen van ingescande foto's verloren gaan. Een tweede nadeel van het ImagePac-formaat is het gebruik van zijn eigen niet-gestandaardiseerde compressiemethode. Een foto van 2048 x 3072 pixels neemt hierdoor slechts tussen 4 en 6 megabytes in beslag. Dezelfde foto ongecomprimeerd opgeslagen als TIFF- bestand is groter dan 18 megabytes. Het ImagePac-formaat werd volledig nieuw gebouwd door Kodak. Kodak baseerde zich zelfs niet op een bestaand bestandsformaat en riep zelfs een nieuw kleurenschema in leven (Photo YCC). ImagePac-afbeeldingen hebben een kleurdiepte tot 24 bits.

Het grote nadeel van dit bestandsformaat is zijn afhankelijkheid van producent Eastman Kodak en de nodige speciale hard-en software. Gezien de hoge kostprijs die hieraan was verbonden, is het bestandsformaat en de technologie nooit volledig bij de gewone consumenten doorgebroken. Begin de jaren 1990 vond het wel een brede ingang in de grafische sector. Grafisch programma's als Photoshop en ACDSee kunnen anno 2002 ImagePac-bestanden wel nog openen, maar de kans is gering dat dit op lange termijn zal mogelijk blijven. Precies vanwege deze redenen heeft het Stadsarchief Antwerpen zijn afbeeldingen op Kodak Photo-CD's inmiddels naar ongecomprimeerde TIFF-bestanden omgezet.

Kodak bracht nadien nog PictureCD als goedkoper en gebruiksvriendelijk alternatief op de markt, maar deze toepassing kende evenmin een groot succes. De JPEG-afbeeldingen op een Picture CD's hebben een vaste resolutie van 1536 x 1024 pixels. De CD's worden conform de ISO-9660 standaard gemaakt. De opkomst van de digitale camera heeft deze technologie volledig op de achtergrond gedwongen.

Referentie: http://www.kodak.com/US/en/digital/dlc/book2/chapter4/imgpacp1.shtml

[35] Deze laatste resolutie wordt enkel gebruikt bij foto's met afmetingen groter dan 5 x 7 inches. Deze resolutie is enkel beschikbaar op een Pro Photo CD master disk.

C.3.2 Vectorafbeeldingen

C.3.2.1 Officiële standaarden

A) CGM: Computer Graphics Metafile

CGM werd in 1987 als officiële ISO-standaard 8632 vastgelegd (*ANSI X3.122(1986): American National Standard for Information Systems - Computer Graphics Metafile for the storage and transfer of picture description information; ISO/IEC-8632(1999): Information technology -- Computer Graphics Metafile for the storage and transfer of picture description information*). In verschillende afzonderlijke landen (o.a. VS, VK) werd CGM als nationale standaard vastgelegd. CGM wordt gebruikt als standaard voor de uitwisseling en de archivering van afbeeldingen. Dit is mogelijk door de hard-en software onafhankelijke manier waarop de CGM-bestanden worden beschreven. Verschillende producenten ontwikkelen toepassingen voor het maken en bekijken van CGM-bestanden.

Er bestaan drie versies van CGM. De eerste versie werd in 1987 als standaard vastgelegd en was ontwikkeld met het oog op gebruik binnen en uitwisseling tussen CAD- en grafische toepassingen. Er werden 90 elementen voorzien. Versie 3 (ISO-8632: 1992) breidde CGM uit met curven en bijkomende grafische attributen voor technische tekeningen (30-tal bijkomende elementen). Versie 4 (ISO-8632:1999) voegde applicatie structurering aan CGM toe. Dit laat de opname van niet-grafische informatie in een CGM-bestand toe (bijv. hyperlinks).

Een CGM-bestand kan in principe twee dimensionele vectorafbeeldingen, rasterafbeeldingen of een combinatie van beide bevatten maar wordt in de praktijk meestal voor statische vectorafbeeldingen gebruikt. Hiervoor was CGM initieel ontworpen. CGM-bestanden bevatten geen animatie of dynamische effecten. CGM dient voor de uitwisseling, het transport en het vastleggen van afbeeldingsbeschrijvende informatie. CGM-bestanden zijn platformonafhankelijk.

CGM-bestanden zijn metabestanden. Ze zijn samengesteld uit elementen. De elementen bevatten de vormen en hun verschijningsvorm. De CGM-standaard bepaalt welke elementen op welke positie in het bestand worden opgeslagen. Voor de toepassing van specifieke CGM-functionaliteiten binnen bepaalde sectoren zijn verschillende CGM-profielen ontwikkeld. Een profiel is een bepaalde interpretatie van de CGM-regels waarbij slechts een beperkt aantal elementen en attributen worden gebruikt. In een profiel worden doorgaans een aantal regels strenger toegepast dan de officiële CGM-standaard voorziet. Voorbeelden van dergelijke profielen zijn PIP (petroleumindustrie), ATA (luchtvaartindustrie) en CALS (DoD). Het toepassen van een bepaald profiel vergemakkelijkt de uitwisseling binnen specifieke toepassingen. Het profiel WebCGM beschrijft hoe CGM-bestanden binnen webbrowsers worden gebruikt. WebCGM is een W3C-Recommendation (1999) en wordt in de CAD-gemeenschap gebruikt voor de presentatie van technische tekeningen.

Er zijn drie verschillende encodings waarin men een CGM-bestand kan bewaren: clear text (best voor editing of programmeren), character (best voor uitwisseling want de kleinste bestandsomvang) en binair (snelst toegankelijk). Voor een systeemonafhankelijke opslag wordt bij voorkeur de character encoding gebruikt.

CGM wordt veel gebruikt binnen grafische databanktoepassingen en voor de uitwisseling van vectorafbeeldingen. Veel grafische computerapplicaties ondersteunen CGM. CGM wordt ook gebruikt binnen Internettoepassingen. CGM is erkend als een afzonderlijke MIME-type. Het DoD nam CGM

aan als standaard. De DoD toepassing van CGM wordt het CALS CGM-profiel genoemd. CGM is in de meeste tekenprogramma's geïmplementeerd. Voor het bekijken van CGM-afbeeldingen in een webbrowser is een plug-in (bijv. WebVIEW CGM) nodig. Met deze plug-in kan een CGM-bestand op dezelfde manier als een JPEG- of GIF-afbeelding worden bekeken. CGM kan bijvoorbeeld gebruikt worden als het archiveringsformaat voor vectorafbeeldingen die in CorelDRAW (*.cdr-bestanden) worden gemaakt. CGM wordt ook gebruikt om technische tekeningen in te bewaren. Dit is onder meer het geval in de ruimte- en luchtvaartsector en de automobielindustrie. Standaard CGM wordt compressieloos toegepast.

Referentie: http://www.cgmopen.org; http://www.iso.ch; http://www.w3.org/TR/REC-WebCGM/

C.3.2.2 Defacto standaarden

A) SVG: Scalable Vector Graphics

De SVG-specificatie (versie 1.0) werd op 4 september 2001 vastgelegd door het *World Wide Web Consortium*. SVG is een toepassing van XML om tweedimensionele vectoriële (eventueel gemengd met raster) afbeeldingen vast te leggen. SVG-afbeeldingen kunnen ook animatie of tekst bevatten en interactief zijn. Momenteel wordt al werk gemaakt van het vastleggen van SVG 1.1. Het geregistreerde MIME-type is image/svg+xml.

SVG heeft dezelfde status als XML: niet producent gebonden, open en vrij, platformonafhankelijk, Recommendation van het *W3C*. Er zijn verschillende softwareapplicaties voor het maken of bekijken van SVG-bestanden. Er bestaan zowel stand alone viewers en editors als plug-ins voor het bekijken van SVG-bestanden in een webbrowser[36] (bijv. de SVG-viewer van Adobe). De SVG-specificatie maakt een onderscheid tussen statische en dynamische viewers. De eerste groep applicaties tonen enkel het SVG-bestand als een statisch document. De dynamische viewers geven toegang tot de interactieve en dynamische componenten van de SVG-afbeelding.

Net zoals XML-bestanden hebben SVG-bestanden een Unicodebasis en kunnen ze bijgevolg door verschillende computerapplicaties worden geopend. Hierdoor hebben SVG-bestanden ook een relatief kleine bestandsomvang. Net zoals alle andere vectoriële afbeeldingen kan SVG zonder kwaliteitsverlies geschaald worden of kan er op geselecteerde gebieden ingezoemd worden. In tegenstelling tot de rasterafbeeldingen kan er op tekst in het SVG-bestand worden gezocht. De tekst kan eveneens geselecteerd worden. SVG-bestanden kunnen eveneens in combinatie met stylesheets worden gebruikt. Naast CSS is er de mogelijkheid om met XSLT de SVG-bestanden te transformeren. In de SVG-syntaxis is een metadata-element voorzien waarin metadata over het archiefdocument kan worden vastgelegd. De metadata worden in het bestand zelf ingebed.

Referentie: http://www.w3.org/TR/SVG/

[36] http://www.w3.org/Graphics/SVG/SVG-Implementations.htm8

B) DXF: Drawing eXchange Format

DXF is het bestandsformaat van producent Autodesk voor de uitwisseling van AutoCADbestanden (*.dwg-bestanden). De DXF-specificatie wordt door Autodesk vrijgegeven.

Er bestaan verschillende versies van het DXF-bestand. De DXF ASCII versie was al geïmplementeerd in AutoCAD 1.0 (december 1982). Deze DXF-versie kan het best vergeleken worden met de ASCII-versie van het binaire en meer compacte DWG-formaat. Vanaf AutoCAD versie 10 is er ook een binaire versie van het DXF-formaat voorzien. De binaire versie neemt in het algemeen 25 % minder schijfruimte in beslag dan de ASCII versie en wordt volgens Autodesk 5 keer zo snel gelezen of geschreven door AutoCAD. De ASCII-versies zijn echter het gemakkelijkst uit te wisselen en het bewaren van CAD-tekeningen als ASCII-DXF-bestanden gaat met minder informatieverlies gepaard dan bij binaire DXF-bestanden. Ook de ASCII en binaire versie hebben verschillende versies. Het DXF-formaat evolueert immers mee met de AutoCADmogelijkheden en het DWG-formaat. De recentste DXF-versie is 16.1.01.

Binnen het DXF-formaat wordt tagging gebruikt om de onderdelen van de tekening te identificeren en te definiëren. Een tag wordt in DXF aangeduid met een geheel getal dat aanduid welk datatype wordt beschreven. De tags in een DXF-bestand zijn groepscodes. Bij elke nieuwe versie worden nieuwe groepcodes of tags opgenomen.

De omzetting naar DXF wordt best wel op informatieverlies gecontroleerd. Of de uitwisseling van tekeningen op basis van DXF-bestanden lukt, is in veel gevallen afhankelijk van de filters die door de export- of importapplicatie wordt gebruikt. De gebruiker heeft de keuze tussen een full DXF-export en een partial export. Bij full export worden alle componenten van een tekening, inclusief blockdefinities, lijntypes, layer informatie, dimensiestijlen, enz. mee opgeslagen. Bij een partial export worden enkel de geselecteerde onderdelen geëxporteerd.

DXF is het formaat op basis waarvan de uitwisseling van CAD-bestanden tussen pakketten zoals Autocad, Microstation en MiniCAD gebeurt. Daarnaast zijn er nog vele andere computerapplicaties die DXF ondersteunen. Alvorens DXF-bestanden binnen een andere applicatie te gebruiken, is het aangewezen om te controleren of wel alle layers worden meegenomen. Andere CAD-applicaties hanteren immers een beperking op het aantal layers.

Referentie: AUTODESK, *DXF Reference Guide*, 2001 (http://www.autodesk.com/techpubs/autocad/dxf/).

C) DWG

DWF is het native bestandsformaat van AutoCAD. Binnen de wereld van CAD/CAM-toepassingen heeft AutoCAD door zijn wijdverspreidheid de status verworven van standaard. DWF is een binair en heel compact bestandsformaat waarvan de specificatie door producent Autodesk niet wordt vrijgegeven. Er bestaan dan ook nagenoeg (of helemaal geen?) andere applicaties dan AutoCAD die *.dwg-bestanden kunnen openen. Om *.dwg-bestanden toch met andere CAD-applicaties te kunnen uitwisselen, heeft Autodesk het DXF-formaat ontworpen. De DXF-specificatie is wel gepubliceerd en

vrij beschikbaar. AutoCAD gebruikte aanvankelijk ook IGES voor de uitwisseling van CAD-bestanden, maar die ondersteuning lijkt momenteel weggevallen.

Bij elke nieuwe wijziging van AutoCAD wordt ook het DWG-formaat aangepast. Dit is volgens Autodesk nodig voor de opslag van nieuwe objecttypes en de implementatie van nieuwe opslagmethoden. Oude AutoCAD-versies kunnen bestanden gemaakt met een nieuwere versie doorgaans niet inlezen. Aan de andere kant is er wel tot op zekere hoogte achterwaartse compatibiliteit.

Door het wijdverspreide gebruik van AutoCAD kan DWG in de praktijk als uitwisselingsformaat worden gebruikt. Voor de archivering van DWG-bestanden is er echter nog steeds geen producent- of versieonafhankelijke oplossing voor handen die niet met functionaliteits- of informatieverlies gepaard gaat. Er is overigens voor geen enkel CAD-formaat een officieel gestandaardiseerd of publiekelijk gepubliceerd archiveringsformaat beschikbaar. In Nederland legt de *Regeling geordende en toegankelijke staat* op dat CAD-tekeningen als PDF-bestanden worden gearchiveerd (art. 6). Het *Center for the study of Architecture/Archaeology* (CSA) houdt een CAD-archief bij waarin tekeningen belangrijk voor archeologie en architectuurgeschiedenis in digitale vorm worden gearchiveerd. Het CSA CAD-archief houdt de tekeningen in hun oorspronkelijk DWG-versie bij tot dat de recenste AutoCADversie niet meer in staat is om bepaalde versies in te lezen. Een andere mogelijkheid is het publiceren van de CAD-tekeningen als EPS-bestanden.

V. ONTSLUITING & TOEGANKELIJKHEID

A. TOPIC MAPS

F. BOUDREZ, *XML Topic Maps voor digitale archivering*, Antwerpen, 2002. (http://www.antwerpen.be/david → cases).

VI. DRAGERS

Voor het fixeren van digitale informatie op een drager wordt voor het ogenblik vooral gebruik gemaakt van ferromagnetisme of van lasertechnologie. Bij ferromagnetisme worden magnetische pigmenten in een laag boven de polethyleen-terephthalaatlaag bewaard. We beschikken over magnetische dragers in de vorm van diskettes en allerhande soorten tapes en cartridges. Bij optische schijven wordt informatie door middel van een laserstraal in de vorm van pits en lands opgeslagen. De pits worden in de polycarbonaatlaag gedrukt (geperste schijven) of in de polymeer pigmentlaag gebrand (CD/DVD-R en CD/DVD-RW). Voor de optische opslag van digitale informatie worden voor het ogenblik hoofdzakelijk schijven gebruikt, maar momenteel wordt ook gewerkt aan een tapeformaat waarop informatie met een lasterstraal wordt bewaard. Ongeacht de gebruikte technologie zijn er twee manieren waarop men data kan bewaren: analoog (lineaire opslag: LP, audiotape, videotape, CD, HD-Rosetta) en digitaal (niet-lineaire opslag: CD, CD-ROM). Algemeen wordt digitale opslag verkozen boven analoge. Het kopiëren van analoge informatie gaat doorgaans met kwaliteitsverlies gepaard, wat bij digitale informatie niet het geval is.

A. MAGNETISCHE DRAGERS

In het algemeen worden magnetische dragers geacht minder aangewezen te zijn voor de opslag van digitale archiefdocumenten. Als reden wordt hiervoor naar hun beperkte levensduur verwezen. Amerikaans onderzoek in de jaren 1990 wees echter uit dat met de huidige technologie ook magnetische dragers een levensduur van 10 tot 30 jaar hebben[37]. De dragers dienen wel goed behandeld te worden en in goede klimatologische omstandigheden bewaard te worden[38].

Magnetische dragers blijven echter kwetsbaarder dan optische dragers: magnetisme neemt af en het contact tussen band en leeskop veroorzaakt slijtage. CD en DVD hebben ook een snellere gegevensoverdracht en -toegang. Magnetische dragers moeten ook regelmatig herspoeld worden. De lage kostprijs per opgeslagen megabyte speelt dan weer in het voordeel van magnetische dragers. Grote buitenlandse archiefinstellingen zoals het NARA of de Nationale Australische archiefdienst maken volop van tapes gebruik voor de opslag van grote digitale archiefbescheiden. Bepaalde tapes hebben een opslagcapaciteit tot 1 terabyte.

Niet alle soorten magnetische dragers zijn geschikt om analoge of digitale informatie te archiveren. Diskettes en harde schijven zijn niet duurzaam genoeg en laten geen systeemonafhankelijke opslag toe. Open haspels raken meer en meer in onbruik. Cassettes en cartridges zijn wel geschikt voor de

---

37 J.W.C. VAN BOGART, *Mag tape life expectancy 10-30 years*, Brief aan Scientific American, 13 maart 1995.

38 Voor meer informatie over het gebruik van magnetische dragers voor de archivering van digitale archiefdocumenten, zie: http://www.antwerpen.be/david → cases → *Magnetische dragers voor het archief*, Antwerpen, 2002.

lange termijnbewaring van archiefdocumenten. De cassettes en cartridges dienen hiervoor wel aan een aantal vereisten te voldoen:

- ☑ de cassette/cartridge is een fysieke standaard: de passende afspeelapparatuur is in de toekomst beschikbaar
- ☑ de informatie is in een logische standaard opgeslagen: het bestandssysteem en de bestandsformaten waarin archiefdocumenten zijn opgeslagen, is leesbaar door toekomstige technologieën
- ☑ na opname: zo hoog mogelijke kwaliteit (afhankelijk van apparatuur en blanco tape) want magnetisme neemt af
- ☑ lange termijn bewaring zonder informatieverlies door goede bewaring en behandeling

A.1 Fysieke standaarden: cassettes en cartridges

Voor de archivering van magnetische banden dient men in de toekomst op zijn minst over de nodige afspeelapparatuur te beschikken. Er zijn vele fysieke standaarden voor cassettes en cartridges vastgelegd. Voor archiveringsdoeleinden is het aangewezen om een magnetische band met de *lineaire* opnamemethode (longitudinal of serpentine) te gebruiken. Het LTO-formaat is een voorbeeld van een defacto standaard, terwijl DLT als een officiële standaard is vastgelegd. De lijst van vastgestelde fysieke standaarden voor cassettes en cartridges is vrij lang en wordt best op de websites van de standaardiseringsinstanties geraadpleegd.

A.2 Logische standaarden: labeled tapes

Voor archiveringsdoeleinden gaat de voorkeur uit naar een labeled tape. Dergelijke tapes bieden het voordeel dat gegevens over de computerbestanden op de tape zelf worden opgeslagen in zogenaamde labels. Het betreft onder meer: bestandsnamen, opnamemethode, blocklengte, recordlengte, data, enz. Bij unlabeled tapes wordt deze informatie niet op de tape zelf opgeslagen, maar moet deze in externe documentatie worden vastgelegd.

A.2.1 ISO-1001

ISO-1001(1986): Information processing -- File structure and labelling of magnetic tapes for information interchange. Deze standaard regelt de bestandenstructuur en het gebruik van labels op tapes die worden gebruikt voor de uitwisseling van informatie. Volgens specialisten bevat deze ISO-standaard echter een aantal fouten, en is het beter om de ANSI-standaard toe te passen. De meeste tapeapplicaties zijn dan ook in de eerste plaats ANSI-label of IBM Standard Label compatibel.

A.2.2 ANSI LABEL X3.27 / ANSI INCITS 27-1987 (R1998)

Net zoals ISO-1001 specificeert deze ANSI-standaard het volume, de bestandenstructuur, de karakteristieken van de blocks en de labels die worden gebruikt voor de identificatie van de records (ANSI INCITS 27-1987 (R1998), *File structure and labeling of magnetic tapes for information interchange*). Deze standaard werd al meermaals aangepast. De oudste versies dateren uit het einde van de jaren 1960. De laatste herziening vond in 1998 plaats. Tapes voldoen aan deze standaard wanneer alle informatie conform de voorschriften op de tape geschreven is.

Een volume bestaat uit een opeenvolging van blocks en tape marks. De blocks bevatten records. Blocks en records hebben een vaste lengte of een variabele lengte. Tape marks zijn controle blocks die als delimeter worden gebruikt. De labels hebben een vaste lengte van 80 bytes en nemen de eerste posities van een block in beslag. Er zijn verplichte en optionele labels. De verplichte labels zijn volume header nr. 1, file headers 1 en 2, file trailers 1 en 2, end of file, end of volume. De optionele labels zijn volume header nr. 2 tem 9, de file headers nr. 3 tem 9, de user file headers nr. 1 tem 9, de file trailers nr. 3 tem 9, de user file trailers nr. 1 tem 9. De optionele labels dienen echter om applicatiegebonden informatie bij te houden en worden bijgevolg beter niet gebruikt.

Van de verplichte labels kan de gebruiker of de archivaris enkel de volume header label invullen. De andere verplichte labels worden door de tapeapplicatie ingevuld.

VELDEN VOLUMELABEL	AANTAL KARAKTERS	INHOUD
label identifier	3	VOL
label number	1	1
volume identifier	6	tape ID
volume accessibility	1	spatie: geen beperking elk ander karakter: wel beperking
vrije posities	13	
implementation identifier	13	software ID
owner identifier	14	ID eigenaar/creator
vrije posities	28	
label standard version	1	standaardversienummer (1, 2, 3 of 4)

De standaard onderscheidt 4 levels van gegevensuitwisseling.

Level 1:	volumeset bestaat uit één file, alle records zijn fixed-length, bestandsnamen zijn beperkt tot 17 karakters
Level 2:	volumeset kan uit meerdere files bestaan, alle records zijn fixed-length, bestandsnamen zijn beperkt tot 17 karakters
Level 3:	volumeset kan uit meerdere files bestaan, alle records zijn ofwel fixed-length ofwel variable-length, bestandsnamen tot 80 karakters
Level 4:	geen beperkingen, bestandsnamen tot 80 karakters

ANSI Labeled tapes zijn doorgaans ASCII-encoded (labels en data).

A.2.3 IBM Standard Label

IBM Standard Label is de defacto standaard die in de eerste plaats in IBM mainframeomgevingen wordt gebruikt. Er zijn een aantal verschillen tussen het ANSI labelformaat en het IBM Standard labelformaat.

	ANSI Label	IBM Standard Label
owner identifier (volumelabel):	14 karakters	10 karakters
lengte bestandsnamen	max. 80 karakters	max. 17 karakters
encoding	ASCII	EBCDIC
header label 3 en 4	optioneel	niet beschikbaar

A.2.4 System Independant Data Format (SIDF)

SIDF is een officiële standaard voor een hard- en software onafhankelijk opslagsysteem voor computergegevens en hun primaire bestandssysteeminformatie zoals data, attributen en karakteristieken. De standaard bepaalt hoe computerdata op een medium logisch worden georganiseerd en wat de vereisten zijn voor configuraties om deze media te kunnen maken en/of in te lezen.

SIDF is hoofdzakelijk ontworpen door Novell op basis van haar Storage Management System (SMS) voor NetWare, een defacto standaard. SIDF werd eerst door de European Computer Manufacturer's Association (ECMA-208, 1994) als standaard vastgelegd en in 1996 door ISO (*ISO-14863: Information technology -- System-Independent Data Format (SIDF)*). Het beheer van de standaard is in handen van de SIDF Association. Leden zijn onder andere Cheyenne Software, Colorado Memory Systems, Dantz Development, Emprise Technologies, Exabyte, Hewlett-Packard, Legato Systems, Mountain Network Solutions, NovaStar Corporation, Novell, Seagate Storage Management, Stac Electronics, Sytron en Symantec. Microsoft maakt geen deel uit van de SIDF Association, want het heeft zijn eigen Microsoft Tape Format. De SIDF Association staat onder meer in voor certificatie van software die aan de standaard voldoet[39].

Een SIDF-drager is fysiek onderverdeeld in sectoren met een gelijke lengte van 2^{n+8} bytes waarbij n een positief geheel getal is (512, 1024, 2048, 4096, ...). Elke fysieke partitie komt overeen met één volume. Een volume bestaat uit een preambule (o.a. header), de dataruimte (met file sets) en een postambule. Het einde van het volume wordt aangegeven door de volume terminator.

SIDF is toepasbaar op elke type drager, zowel magnetisch als optisch. In de praktijk wordt SIDF hoofdzakelijk voor twee doeleinden gebruikt: de gezamenlijke bewaring van data gegenereerd op computers met verschillende besturingssystemen (bijv. DOS, Unix, OS/2, FTAM) op één en dezelfde server en voor het maken van platformonafhankelijke backups. Bij het wegschrijven of backuppen van computerbestanden op schijven of tape wordt immers hoofdzakelijk een typisch apparaat en software gebonden bestandssysteem gebruikt waardoor de tapes niet uitwisselbaar zijn. De schijven of tapes

[39] http://www.cs.wisc.edu/~jgast/sidf

zijn hierdoor gebonden aan één specifiek besturingssysteem(versie). Binnen het SDIF-bestandssysteem worden blocks systeem- en bestandsdata door tags gedocumenteerd. In de tags of field identifiers worden de attributen van de bestanden opgeslagen. Op basis van deze tags wordt het type informatie in de blocks gecommuniceerd naar elk SIDF compatibel systeem. Hierdoor wordt de uitwisseling tussen verschillende (versies van) besturingssystemen en backupsoftware mogelijk gemaakt. Door de toepassing van SDIF is het mogelijk om bestanden geschreven door een bepaald besturingssysteem te laten inlezen door een computer met een ander besturingssysteem.

Aanvankelijk was SIDF in de eerste plaats in NetWarebackuptoepassingen geïmplementeerd (bijv. SBACKUP), maar momenteel zijn er voor de andere courante besturingssystemen SIDF compatibele backupapplicaties beschikbaar. Een aantal toepassingen zijn wel beperkt tot het inlezen van SIDF compatibele tapes, en kunnen zelf geen SIDF-tapes schrijven.

SDIF is ook uitbreidbaar, waardoor het kan worden aangepast aan nieuwe functionaliteiten van toekomstige (besturings)systemen. SIDF kan zowel gecomprimeerd als ongecomprimeerd worden toegepast.

Referentie: http://www.cs.wisc.edu/~jgast/sidf/

B. OPTISCHE DRAGERS

B.1 Compact Disk (CD en CD-ROM)

Voor CD's worden er twee soorten standaarden gehanteerd[40]. De fysieke standaarden bepalen de structuur, de ordening en het gebruik van bytes op de CD. De audioCD's zijn gebaseerd op de red book specificatie (IEC-908). Voor gegevensCD's is de officiële standaard ISO-10149, de defacto standaard is CD-ROM XA (eXtended Architecture). De logische standaarden hebben betrekking op het bestandssysteem van de CD. In tegenstelling tot magnetische dragers gebruiken CD's immers een ander bestandssysteem dan de computerbesturingssystemen. De officiële standaard is hiervoor ISO-9660. CD's beschreven conform deze standaard, zijn platformonafhankelijk en uitwisselbaar. Rock Ridge en Joliet zijn op ISO-9660 gebaseerd en tot op zekere hoogte uitwisselbaar. Beide bestandssystemen zijn defacto standaarden. De uitzondering hierop is het Hierarchical File System (HFS) van Macintosh. HFS is niet gebaseerd op ISO-9660 en kan niet als een standaard worden beschouwd.

B.1.1 Fysieke standaarden

B.1.1.1 Officiële standaard

A) AudioCD's: IEC-908

[40] Een meer uitgebreide beschrijving van de CD-standaarden vindt u in het artikel *CD's voor het archief.* (zie: http://www.antwerpen.be/david : publicaties ? overige publicaties)

AudioCD's worden beschreven conform de standaard IEC-908. Bij het maken van audioCD's wordt automatisch deze standaard toegepast. AudioCD's zijn perfect uitwisselbaar en platformonafhankelijk. Een audioCD wordt in één sessie geschreven en bestaat uit 3 delen: de lead-in, de audiotracks en de lead-out. Een audioCD kan maximaal 99 tracks of ongeveer 74 minuten geluid bevatten. De foutopsporings- en verbeteringscode voor audioCD's blijft beperkt tot CIRC. Hierdoor kan er per sector voor 2352 bytes aan gebruikersdata worden gebruikt.

B) GegevensCD: ISO-10149

ISO/IEC 10149(1995): Information technology -- Data interchange on read-only 120 mm optical data disks (CD-ROM). De ISO-10149 standaard biedt voor het beschrijven van gegevensCD's twee modes aan. CD's met computerbestanden of programma's (CD-ROM's) worden gecreëerd op basis van mode 1. Mode 1 gebruikt naast CIRC nog een tweede foutopsporings- en verbeteringsschema: EDC/ECC. Per sector wordt een deel voor EDC/ECC in beslag genomen, waardoor er maar 2048 bytes voor gebruikersdata beschikbaar blijft. Ten slotte is er ook nog mode 2. Mode 2 hanteert opnieuw enkel CIRC voor de opsporing en verbetering van fouten, maar wordt in de praktijk nauwelijks gebruikt. In de plaats van mode 2 wordt CD-ROM XA gebruikt. Goede software voor het schrijven van CD's laat de gebruiker toe te kiezen tussen ISO-10149 (doorgaans aangeduid als CD-Rom mode 1) en CD ROM XA.

B.1.1.2 Defacto standaard

A) CD-ROM XA

CD-ROM XA (eXtended Architecture) is een uitbreiding van ISO-10149 door Sony, Philips en Microsoft en dient voor multimediacd's (mengeling van tekst, muziek, video, afbeeldingen, …) waarbij de modes 0 en 1 door elkaar worden gebruikt. CD-ROM XA bestaat uit mode 2/form 1 (voor computerbestanden en programma's; 2048 bytes gebruikersdata per sector) en mode 2/form2 (voor audio en video; 2324 bytes gebruikersdata per sector). CD-ROM XA wordt ook gebruikt bij Kodak Photo-CD. CD-ROM XA laat ook multisessies toe. Bij ISO-10149 is dit in theorie ook mogelijk, maar levert in de praktijk veelal problemen op zodat dit best wordt vermeden[41].

[41] http://www.roxio.com/en/support/cdr/multisessionhistory.html

B.1.2 Logische standaarden: Bestandssystemen

B.1.2.1 Officiële standaard

A) ISO-9660

De officiële benaming van de standaard is: *ISO-9660(1988): Information Processing -- Volume and file structure of CD-ROM for information interchange*. De laatste wijziging dateert van 1995. De bedoeling was een standaard te creëren die inleesbaar was op elke computer, onafhankelijk van het bitpatroon, het computerbesturingssysteem of de applicatie waarbinnen data worden opgevraagd. Als gevolg daarvan werden in de ISO-norm een aantal beperkingen opgenomen zodat CD-ROM's ook door de zwakkere besturingssystemen konden worden ingelezen.

Voor de gebruiker is het bestandssysteem belangrijk voor de mappenstructuur, de map- en bestandsnamen en de volume descriptor op de CD.

ISO-9660 onderscheidt drie levels voor de uitwisseling van CD's. Level 1 kent de meeste beperkingen en is grotendeels aan het bestandssysteem van MS-DOS ontleend. Level 2 en 3 zijn uitbreidingen hierop, maar worden niet meer ondersteund door computers met MS-DOS als besturingssysteem. Deze levels worden wel ondersteund door meer geavanceerde besturingssystemen zoals Unix, Windows (FAT-32, NTFS), Novell, Macintosh, enz[42].

Level 1:
- ✓ mappenstructuur: max. 8 niveau's, de rootmap inbegrepen
- ✓ mapnaam: max. 31 karakters, extensies zijn niet toegelaten, hoofdletters
- ✓ pathlengte: max. 64 karakters (8 niveau's x 8 karakters)
- ✓ volumenaam: 11 alfanummerieke karakters, hoofdletters
- ✓ bestandsnaam: max. 8 karakters en 3 karakters voor de extensie. Bestandsnaam en extensie worden van elkaar gescheiden door een punt. Alle karakters staan in hoofdletters. De toegestane karakters zijn A-Z, 0-9, een punt en het underscoreteken. De karakters ! " % & ' ( ) * + - . / :; < = > ?, de komma of een spatie mogen niet worden gebruikt.
- ✓ ordening van de bestanden: één bestand wordt aaneensluitend op de CD geplaatst (geen inter-leaving). De bestanden worden teruggevonden op basis van hun beginpositie en lengte.

Level 2: Level 2 laat ten eerste langere map- en bestandsnamen toe: max. 32 karakters en 3 karakters voor de extensie. Bestandsnaam en extensie worden van elkaar gescheiden door een punt. Ten tweede kunnen de karakters zowel hoofdletters als kleine letters zijn. De toegestane karakters zijn bijgevolg: A-Z, a-z, 0-9, '.' en '_'. De karakters ! " % & ` ( ) * + - . / :; < = > ?, de komma of een spatie mogen niet worden gebruikt. Level 2 hanteert hetzelfde ordeningssysteem als level 1.

Level 3: Level 3 hanteert dezelfde map- en bestandsnamen als level 2. Het verschil is echter dat de bestanden in level 3 niet aaneensluitend moeten zijn (wel inter-leaving). Hierdoor kunnen de bestanden als pakketten op de CD worden geplaatst.

---

[42] ISO-9660: 1988, Information processing - Volume and file structure of CD-ROM for information interchange. De beschrijving van de standaard is beschikbaar op: http://www.y-adagio.com/public/standards/iso_cdromr/. Zie ook: F. CAFFARELLI en D. STRAUGHAN, *Publish yourself on CD-ROM*, p. 73-114.

In Nederland, Denemarken en bij de NASA bijvoorbeeld wordt ISO-9660 als standaard gebruikt. De CD's die bij het archief worden neergelegd, moeten voldoen aan deze norm. Het stadsarchief Antwerpen past eveneens ISO-9660 als standaard toe[43]. Bij het opleggen van ISO-9660 als norm wordt best ook exact voorgeschreven welk level van ISO-9660 wordt toegepast.

B.1.2.2 Defacto standaarden

A) Rock Ridge

Rock Ridge is de benaming voor het CD bestandssysteem van Unix. Rock Ridge laat langere bestandsnamen toe en de mappenstructuur kan dieper zijn dan 8 niveau's. De andere uitbreidingen maken het mogelijk dat voor elke map of bestand gebruikersrechten en tijdstippen worden bijgehouden. Rock Ridge laat ook symbolic links toe. Rock Ridge wordt enkel door Unix ondersteund. Er is een uitbreiding van Rock Ridge voor Amiga computers[44].

B) Joliet

De uitbreiding op ISO-9660 voor Windowscomputers wordt Joliet genoemd. Joliet wordt ondersteund voor (bijna) alle computerprogramma's die op Windows lopen. De software voor de creatie van CD's onder Windows biedt doorgaans de mogelijkheid om een CD te maken op basis van ISO-9660 of van Joliet[45]. De uitbreidingen zijn mogelijk omdat Joliet naast de primary volume descriptor ook gebruik maakt van de supplementary volume descriptor.

De uitbreidingen van Joliet zijn:

- ✓ bestandsnamen tot 64 karakters
- ✓ Unicode codetabel voor map- en bestandsnamen
- ✓ mapnamen kunnen extensies bevatten
- ✓ geen beperking in de diepte van de mappenstructuur

Joliet map- en bestandsnamen kunnen door de meeste besturingssystemen worden ingelezen. Bij DOS, Macintosh en Unix wordt de bestandsnaam wel afgekort (bijv. BESTAN~1.txt ipv bestandsnaam.txt). Joliet bewaart immers voor elk bestand samen met de lange bestandsnaam een geassocieerde DOS-bestandsnaam.

---

[43] Nederland: *Regeling duurzaamheid archiefbescheiden*, art. 7; Denemarken: Omzendbrief nr. 4 (14 jan. 2000): *Circular on national authorities' delivery of digital records systems to the Danish State Archives*, § 3; Stadsarchief Antwerpen, *Duurzame CD's. Digitaal Archief: Richtlijn 2*, jan. 2002.

[44] A.S. TANENBAUM, *Modern Operating Systems*, New Jersey, 2001, p. 430-435; F. CAFFARELLI en D. STRAUGHAN, *Publish yourself on CD-ROM*, p. 112-114.

[45] http://bmrc.berkeley.edu/people/chaffee/jolspec.html;

B.2 DVD

DVD stond aanvankelijk voor Digital Video Disk, maar wordt nu als afkorting voor Digital Versatile Disk gebruikt. DVD is de gedoodverfde opvolger van de CD voor muziek en de VHS-videocassetten of de laser disk voor films. In vergelijking met zijn voorgangers kan een DVD-schijf veel meer data (tot 50 gigabyte) bevatten en is de data sneller toegankelijk. De grotere opslagcapaciteit van DVD's is een gevolg van de kortere en fijnere golflengte van de laterstraal (kleinere pits), smallere tracks, een meer efficiëntere kanaalcode en sterkere foutverbeteringscode (RS-PC). In tegenstelling tot CD's kunnen DVD's ook twee lagen bevatten.

Het standaardisatieproces voor DVD is nog volop aan de gang. Onder druk van de computer- en filmindustrie wordt er gewerkt aan standaardisatie, maar dit proces verloopt traag. Aanvankelijk was het zelfs nog geen uitgemaakte zaak welke technologie zou worden gebruikt. Matsushita (o.a. Panasonic en Technics) schoof samen met Time-Warner, Pioneer, Hitachi, e.a. de Super Density Disc (2 zijden, 1 laag) naar voor. Sony en Philips probeerden hun Multimedia CD (MMCD: 1 zijde, twee lagen) door te drukken. Beide technologieën verschillen fundamenteel en zijn incompatibel. In september 1995 kwam er tussen Philips/Sony en de Super Density Disk alliantie een compromis tot stand. Het compromisformaat kreeg de naam DVD mee. In december 1995 werd een DVD Consortium (inmiddels: DVD Forum) in het leven geroepen. Ondertussen beschikken we sinds kort over zowel fysieke als logische standaarden voor DVD-schijven.

DVD is de benaming voor de groep gerelateerde standaarden. Net zoals bij CD's bestaan er verschillende 'boeken' waarin de standaarden zijn vastgelegd. Bij CD's worden deze boeken met kleuren aangeduid, bij DVD's is dat met de letters A tem E.

Er bestaan vijf verschillende toepassingen voor DVD: DVD-ROM: data opslag; DVD-VIDEO: film; DVD-AUDIO: audiotracks; DVD-RECORDABLE: WORM-schijf; DVD-REWRITABLE: herschrijfbare DVD. DVD-ROM, DVD-AUDIO (1999) en DVD-VIDEO (1996) zijn momenteel al beschikbaar. DVD-AUDIO en DVD-VIDEO gebruiken één gemeenschappelijk bestandssysteem: het Universal Disk Format. DVD-ROM's moeten toegankelijk zijn vanop computers en vragen dus compatibiliteit met de gangbare computerbesturingssystemen. DVD-ROM gebruikt het UDF Bridge bestandssysteem. Dit is een combinatie van UDF en ISO-9660. De voornaamste eigenschappen zijn: Unicode-karakters voor bestands- en mapnamen (geen beperkingen) en 256 karakters voor één bestands- of mapnaam. Voor DVD-AUDIO kan PCM (Pulse Code Modulation), Dolby AC-3 Digital en MPEG-audio worden gebruikt. MPEG-1 en MPEG-2 kan toegepast worden bij DVD-VIDEO.

B.2.1 Officiële standaard

B.2.1.1 ISO-13346

ISO-13346: Volume and file structure for write-once and rewritable media using non-sequential recording for information interchange wordt soms ook wel de NSR- of ECMA-167 standaard genoemd. Terwijl ISO-9660 zich beperkte tot een alleen-lezen bestandssysteem is ISO-13346 ontworpen als lezen en schrijven bestandssysteem voor niet-sequentiële opname. Het naleven van de

standaard ISO-13346 brengt platformonafhankelijkheid en vendoronafhankelijkheid met zich mee. ISO-13346 is vrij complex, maar wordt toch vrij algemeen nageleefd[46].

B.2.2 Defacto standaarden

B.2.2.1 Universal Disk Format

Het Universal Disk Format (UDF) is gedefinieerd door OSTA (Optical Storage Technology Association, htpp://www.osta.org). UDF is een subset van ISO-13346. Met UDF wou OSTA een minder complex bestandssysteem realiseren en meer inspelen op de bestaande besturingssystemen. Zo schrijft UDF exact voor hoe en waar een bepaald besturingssysteem welke data op een drager plaatst en hoe een ander besturingssysteem die data verwerkt. UDF wordt dan ook in de plaats van ISO-13346 gebruikt. Net zoals ISO-13346 is UDF bedoeld als platform- en vendoronafhankelijk bestandssysteem voor (herschrijfbare) DVD's. Er bestaan verschillende versies van UDF. De versies 1.02, 1.5 en 2.0 zijn het meest verspreid[47].

UDF heeft de volgende kenmerken:

- max. bestandsomvang: 128 Terabytes
- max. volumeomvang: 128 Terabytes
- map- en bestandsnamen: max. 255 karakters
- karakterset map- en bestandsnamen: Unicode

Voor de overgang van ISO-9660 naar UDF is een 'UDF-bridge' ontworpen. Toepassing van 'UDF-bridge' maakt het mogelijk dat zowel ISO-9660 compatibele lezers als UDF-compabile lezers toegang tot de informatie hebben.

Referentie: *UDF White Paper*

B.2.2.2 Blu-ray disk

Op 19 februari 2002 maakten negen belangrijke technologiereuzen bekend dat ze een gezamenlijke specificatie voor een DVD-recorder vastleggen[48]. De negen betrokken ondernemingen zijn Hitachi, LG Electronics, Matsushita, Philips, Pioneer, Sharp, Samsung, Sony en Thompson Multimedia. Deze standaard wordt de "Blu-ray disc" genoemd. Deze technologie gebruikt een 405 nanomillimeter blauw-violet laserstraal, waardoor er meer informatie op dezelfde oppervlakte kan worden opgeslagen.

---

46 Mededeling prof. dr. Kees van der Meer (E-mail van 7 december 2001).

47 OSTA, *Data interchange & Optical Standards*, Santa Barbara, 1996.

48 http:://www.matsushita.co.jp/corp/news/official.data/data.dir/en020219-4/en020219-4.html

Het vastleggen van deze standaard mag als een doorbraak in de DVD-technologie worden beschouwd. De "Blu-ray disc"-specificatie is een hele vooruitgang in vergelijking met de eerste generatie DVD-recorders. Deze toestellen leveren geen uitwisselbare DVD's af. Hun DVD-schijven kunnen enkel op apparaten van dezelfde fabrikant of groep worden afgespeeld. Deze DVD-apparaten gebruiken een grotere en bredere rode laserstraal waardoor ze standaard *slechts* 4,7 gigabyte kunnen bevatten.

Met deze technologie is het opnemen en herschrijven van uitwisselbare DVD-schijven mogelijk. Op een DVD-schijf met één laag kan tot 27 gigabytes aan digitale informatie worden bewaard. Dit komt overeen met een 2 uur durende digitale hoge definitie film of 13 uur bewegende beelden in VHS-kwaliteit. Deze DVD-schijven zijn dus ideaal voor de opslag van geluid en bewegende beelden in een hoge kwaliteit. Het spreekt voor zich dat deze schijven ook grote hoeveelheden tekstuele informatie kunnen bevatten. Het is de algemene verwachting dat er net zoals bij de huidige DVD-toepassingen ook dubbele lagen DVD's worden gemaakt. Eén DVD-schijf zou uiteindelijk meer dan 50 gigbytes kunnen bevatten.

Typerend voor de blu-raytechnologie is het gebruik van de MPEG-2 compressiemethode voor datatransfer. Hierdoor is een datatransport van 36 megabytes per seconde mogelijk. De beelden die men met een digitale camera registereert kunnen rechtstreeks op een DVD worden geplaatst.

Het vastleggen van de specificatie is nog niet helemaal voltooid. Men verwacht dit tegen de lente van 2002 te kunnen afwerken zodat dan al met de verkoop van de licensies kan worden gestart. De eerste commerciële produkten die de "blu-ray"-technologie gebruiken, mogen in 2003 in de winkels worden verwacht. Of de "blu-ray" DVD-apparaten compatibel zullen zijn met de huidge rode laser DVD's is volledig afhankelijk van de producenten.

C. High-Density Rosetta Rom (HD-ROM)

De HD-ROM is voor het ogenblik zeker nog geen standaard, ook nog geen defacto standaard. Toch is de HD-ROM hier het vermeldenswaard, want deze technologie is speciaal ontworpen om te voldoen aan de archivistische doeleinden. De HD-ROM probeert een antwoord te bieden op de klassieke problemen waarmee elke drager van digitale informatie te maken heeft: beperkte levensduur van de materialen waaruit het medium is samengesteld, zwakheid of gevoeligheid van de gebruikte technologie (bijv. magnetisme), afhankelijkheid van hard- en software om de computerbestanden opnieuw te kunnen inlezen, enz.

De HD-ROM werd ontwikkeld door Los Alamos National Laboratories en wordt verspreid door Norsam Technologies. De HD-ROMtechnologie is gebaseerd op ionstralen. De wavelengte van deze ionstralen bedraagt slechts 25 nm, waardoor de lineaire en trackdichtheid veel groter is, met een veel grotere opslagcapaciteit als gevolg (ongeveer 165 gigabytes). De drager kan uit verschillende materialen bestaan: plastic polymeren, glas, metaal, enz. Norsam gebruikt nikkel als drager. Proeven wezen uit dat deze schijven bestand zijn tegen een temperatuur van 500° Celsius en tegen water. De levensduur van een HD-ROMschijf wordt op 1000 jaar geschat. De informatie wordt analoog opgeslagen zodat de informatie eventueel zonder tussenkomst van hard- en software gewoon met een microscoop kan worden gelezen. Norsam ontwikkelt ook de nodige software om de HD-ROMschijven te bevragen en te raadplegen.

De HD-ROM heeft de ambitie om de drager te zijn voor archiefdocumenten met een permanente levensduur. Bij opslag worden digitale archiefdocumenten omgezet in microns en als zodanig pixel per pixel in de schijf geëtst.

Referentie: http://www.norsam.com/rosetta.html