

Standaarden voor digitale archiefdocumenten

Filip Boudrez
Expertisecentrum DAVID
Antwerpen, 2005

0. INHOUDSOPGAVE

1. BELANG VAN STANDAARDEN VOOR DIGITALE ARCHIVERING.....	1
2. HIËRARCHIE VAN DE ARCHIVERINGSSTANDAARDEN.....	2
3. GESCHIKTE ARCHIVERINGSFORMATEN.....	3
4. CODETABELLEN.....	3
4.1. OFFICIËLE STANDAARDEN.....	4
4.2. DEFACITO STANDAARDEN.....	6
5. BESTANDSFORMATEN.....	7
5.1 TEKSTDOCUMENTEN.....	8
5.2 AFBEELDINGEN.....	22
5.3 AUDIO.....	36
5.4 VIDEO.....	41
5.5 GEOGRAFISCHE INFORMATIE.....	46
6. UITGEBREIDE INHOUDSOPGAVE.....	51

1. BELANG VAN STANDAARDEN VOOR DIGITALE ARCHIVERING

Standaarden zijn belangrijk voor het verzekeren van de leesbaarheid van digitale archiefdocumenten op lange termijn. De computertechnologie evolueert voortdurend met als gevolg dat hard- en software snel verouderen of in onbruik raken. Digitale archiefdocumenten kunnen echter een (middel)lange of zelfs permanente bewaartermijn, en bijgevolg een langere levensduur dan hun oorspronkelijke hard- en softwareomgeving, hebben.

Er bestaan verschillende oplossingen voor het lange termijn leesbaarheidsprobleem¹. Migratie van de archiefdocumenten en emulatie van de nodige hard- en softwareomgeving zijn de meest geciteerde digitale bewaarstrategieën. Standaarden spelen een voorname rol in beide bewaarstrategieën. Bij migratie worden de archiefdocumenten omgezet naar een geschikt archiveringsformaat met de status van standaard. Emulatie is vooralsnog een overwegend theoretische oplossing, maar heeft het meeste kans op slagen wanneer de archiefdocumenten in een gestandaardiseerd bestandsformaat zijn opgeslagen.

Standaarden zijn overigens niet alleen belangrijk voor archiefbeherende instellingen, ook de archiefvormer heeft alle belang bij het toepassen van standaarden. Door standaarden toe te passen vermijdt men dat digitale archiefdocumenten afhankelijk zijn van de omgeving waarbinnen ze werden gecreëerd en dat ze moeten omgezet worden telkens een computerapplicatie in onbruik raakt. Door onmiddellijk de digitale documenten in een gestandaardiseerd formaat te bewaren, vermijdt men ook dat digitale archiefdocumenten eerst omgezet worden, alvorens men ze naar het digitaal archief overbrengt. Tenslotte biedt het toepassen van standaarden het voordeel dat digitale documenten op

¹ Zie hierover F. BOUDREZ, *B. Bewaarstrategieën*, in: F. BOUDREZ EN H. DEKEYSER, *Digitaal archiefbeheer in de praktijk. Een handboek*, Antwerpen-Leuven, 2004.

een uitwisselbare wijze worden opgeslagen. Uitwisselbaarheid houdt in dat andere computers de bits en bytes op de archiefdrager correct kunnen inlezen, verwerken en als een menselijk leesbaar document op scherm presenteren. De uitwisseling van digitale informatie is overigens in de meeste gevallen de voornaamste reden voor standaardisatie.

Naast de belangrijke voordelen voor archiefvormer en archieven, zijn er ook enkele kanttekeningen bij standaarden. 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. Dit biedt dan weer het voordeel dat hierdoor enige stabiliteit is gewaarborgd. Officiële standaardisatie is ten tweede niet altijd gewenst door de ontwerpende ondernemingen want hierdoor verliezen ze voor een stuk hun greep op hun product. Ten derde breiden softwareproducenten vanwege commerciële redenen standaarden soms uit met extra functionaliteiten of implementeren ze standaarden op een applicatiespecifieke wijze. Dit gaat ten koste van de uitwisselbaarheid en is een belangrijk aandachtspunt voor zowel archiefvormer als archiefbeherende instelling. 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 op een hard- en softwarespecifieke wijze documenten te archiveren.

2. HIËRARCHIE VAN DE ARCHIVERINGSSTANDAARDEN

In onderstaand overzicht met technische standaarden worden de standaarden per type archiefdocument hiërarchisch ingedeeld. De hiërarchie is gebaseerd op de officiële status van de standaard. De groep niet-officiële of defacto standaarden is verder onderverdeeld in functie van het beheer en de openbaarheid van de formaatspecificatie. We onderscheiden volgende standaarden²:

- officiële of de jure standaarden: deze standaarden zijn vastgelegd door officiële standaardiseringsorganisaties. Deze internationale, regionale of nationale organisaties danken hun officiële status door de participatie van een (inter-)gouvernementele instanties (bijv. ISO, IEC, ITU, CEN, ITSI, enz.)
- defacto standaarden: deze standaarden zijn vastgelegd door niet-officiële standaardiseringsorganisaties of zijn vanwege hun wijdverspreidheid de norm geworden. De groep defacto standaarden kan nog verder onderverdeeld worden in:
 - open specificaties of aanbevelingen: vastgelegd door niet-officiële normerende organisaties (bijv. W3C, AIIIM, OASIS, IETF), gedocumenteerd. Deze forums, verenigingen of consortia groeperen zich om een standaard uit te werken zonder de nadelen die verbonden zijn aan de officiële standaardiseringsprocedures.
 - open standaarden: afhankelijk van één of meerdere producenten, gedocumenteerd, groot marktaandeel
 - gesloten standaarden: afhankelijk van één of meerdere producenten, niet gedocumenteerd, groot marktaandeel

De officiële standaarden en de open specificaties of aanbevelingen genieten de voorkeur. De technische documentatie over deze standaarden behoort tot het publiek domein. De ontwikkeling en het beheer van die standaarden is in handen van een normerende organisatie die hierbij aan procedures gebonden is. Samen met het feit dat meerdere partijen (publieke instellingen, softwareproducenten, universiteiten, consumenten) bij deze initiatieven zijn betrokken, wordt hierdoor een stukje stabiliteit gewaarborgd.

De open en gesloten defacto standaarden zijn afhankelijk van één producent en kunnen dus ten allen tijde worden gewijzigd. Open defacto standaarden zijn in tegenstelling tot de gesloten standaarden gedocumenteerd. Net zoals bij officiële standaarden of open specificaties is hun samenstelling beschikbaar voor het publiek. In de meeste gevallen kan de samenstelling van het formaat gewoon via de website van de producent gedownload worden en is er zelfs een SDK (software development kit) beschikbaar. Dit is een niet onbelangrijk gegeven, want op basis van deze documentatie is het in

² 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.

principe altijd mogelijk om een nieuwe viewer te creëren. Een tweede voordeel is dat op basis van de specificatie kan gecontroleerd worden of een computerbestand wel aan de formele regels van de standaard beantwoord. Bij gesloten defacto standaarden is dit alles niet mogelijk en deze worden best zoveel mogelijk als archiveringsformaat vermeden.

3. GESCHIKTE ARCHIVERINGSFORMATEN

Een hiërarchische indeling is van belang om het overzicht te behouden en kan als leidraad dienen bij de keuze van een duurzaam archiveringsformaat. De status van standaard en meer bepaald de plaats in de hiërarchie is het eerste criterium voor een geschikt archiveringsformaat. Aan de andere kant is de plaats van een standaard binnen de hiërarchie geen absolute garantie voor de lange termijnleesbaarheid. Enige nuancering is dus op zijn plaats. Voor de leesbaarheid op lange termijn biedt een officiële standaard a priori niet meer garanties dan een open specificatie of aanbeveling. Officiële standaardisatie en marktevoluties lopen immers 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 (defacto standaard) versus SGML (officiële standaard) of van Unicode (defacto standaard) versus ISO-10646 (officiële standaard) illustreert dit.

Naast de standaardisatiestatus gelden bijgevolg nog andere criteria bij de keuze van een geschikt archiveringsformaat:

- de wijdverspreidheid en de marktpenetratie
- de uitwisselbaarheid: de onafhankelijkheid ten opzichte van bepaalde besturingssystemen, netwerkprotocollen en computerapplicaties
- de aanwezigheid van een robuust foutopsporings- en verbeteringsmechanisme: fouten in bitopslag worden automatisch gedetecteerd en zijn herstelbaar
- de mogelijkheid tot systematische en geautomatiseerde validatie: controle of het digitaal archiefdocument wel volledig conform de formaatspecificatie werd opgeslagen
- een goed gestructureerde opslag van informatie
- een opslag zonder informatieverlies: bijv. geen automatische toepassing van compressie
- de mogelijkheid tot insluiten van (zelfgedefinieerde) metadata velden
- de capaciteit om de essentiële eigenschappen van het archiefdocument in tijd over te brengen
- het bewaren van de integriteit van archiefdocumenten
- de mate van autonomie en zelfvoorzienigheid
- de mogelijkheid om op een drager- en apparaatonafhankelijke wijze archiefdocumenten te bewaren
- de eenvoudigheid en gebruiksvriendelijkheid.

4. CODETABELLEN

In codetabellen wordt vastgelegd welke alfanumerieke karakters, leestekens en controle tekens met een bepaalde (hexa)decimale waarde overeenstemmen. Computers verwerken en bewaren immers binaire getallen. Voor de correcte omzetting van de binaire waarden naar menselijk leesbare tekst dient men de overeenstemmende codetabel te gebruiken. Bij de archivering van tekstuele documenten is het van belang om een gestandaardiseerde codetabel als basis te nemen en om heel duidelijk te documenteren welke codetabel werd gebruikt.

4.1. Officiële standaarden

4.1.1 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 afdrubare 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 gemaakt³. De laatste wijziging dateert van 1991.

4.1.2 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 verwerkt⁴. Elke codetabel van de ISO-8859 heeft dezelfde opbouw: de posities 0-127 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.	ISO-8859-12:	/

³ 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 ! " % & ' () * + - . / : ; < = > ?.

⁴ In Latin 1-versie van de ISO 8859 tabel worden de posities 128-159 voor niet-afdrubare controletekens gebruikt. In de karakterset die Windows gebruikt (WinLatin, Windows code page 1252) worden een aantal van deze posities toch voor afdrubare 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.

ISO-8859-5:	Vervangen door ISO-8859-10 Cyrilisch: Russisch, Bulgaars, Servisch, Macedonisch, Oekraïens	ISO-8859-13:	Baltisch: Lets Latin 7
ISO-8859-6:	Arabisch	ISO-8859-14:	Keltisch: Gaelic en Welsh Latin 8
ISO-8859-7:	Grieks	ISO-8859-15:	West-Europese talen met o.a. het euro-teken Latin 9
ISO-8859-8:	Hebreeuws en Jiddisch		Latin 0

4.1.3 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 heeft elk karakter uit elke taal slechts één unieke numerieke waarde. 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 *Unicodeconsortium* 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. Momenteel is Unicode 4.1.0 de laatste versie. De recentste versie van ISO-10646 dateert van 2003.

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. Het Unicodeconsortium wil hiermee bereiken dat de codetabel 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 of 1 byte) naar Unicode (BMP, 16 bits of 2 bytes) verdubbelt de omvang van de bestandsgrootte.

Referentie: <http://www.unicode.org>; <http://www.iso.ch>

4.1.4 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

4.2. Defacto standaarden

4.2.1 UNICODE

Zie 4.1.3 ISO-10646 en Unicode

4.2.2 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 EBCDIC-karakters.

4.2.3 BASE64

Base64 is een encodingswijze waarbij binaire bitstreams worden omgezet naar een bitstream die enkel uit tekstkarakters bestaat. Het Base64-mechanisme is een onderdeel van het gestandaardiseerd internetprotocol MIME RFC 2045. RFC 2045 werd in 1996 vastgelegd door de Internet Engineering Task Force (IETF, 1996) en is een uitbreiding van RFC 822 (IETF, 1982) met de bedoeling om via mail ook de uitwisseling mogelijk te maken van: berichten of headers in een andere encoding dan ASCII, niet-tekstuele en/of samengestelde berichten. RFC 822 beperkte zich immers tot pure tekstmails in ASCII encoding (7 bits). Met Base64 is het mogelijk om via mail binaire bestanden (8 bits) uit te wisselen.

Het Base64 encodings- en decodingsalgoritme is relatief eenvoudig. De originele bitstream wordt in pakketjes van 24 bits (drie octetten) ingelezen. Dit pakket wordt herverdeeld in 4 groepen van 6 bits. De binaire waarde van elk groepje van 6 bits wordt vervolgens gemapt aan het overeenstemmende tekstkarakter in de Base64-tabel. De drie originele octetten worden op die manier vervangen door 4 tekstkarakters. Dit heeft voor gevolg dat de geëncodeerde bitstream ongeveer 1/3 groter is dan de originele bitstream en dus meer bestandsomvang in beslag neemt. Het resultaat is een reeks tekstkarakters die niet menselijk interpreteerbaar is. De outputstream bestaat uit lijnen die maximaal 76 karakters bevatten. Bij encoding wordt verondersteld dat de bitstream geordend is van meest naar minst betekenisvolle bit. Bij decoding wordt het omzettingalgoritme omgekeerd. RFC 2045 schrijft voor dat bij decoding alle karakters die niet voorkomen in de Base64-tabel worden genegeerd. De behandeling van regeleinden vraagt bijzondere aandacht bij encoding en decoding, want fouten zijn veelal een gevolg van de omzettingen van regeleinden.

Tabel 2: De Base64 omzettingstabel

0	A	17	R	34	i	51	z
1	B	18	S	35	j	52	0
2	C	19	T	36	k	53	1

3	D	20	U	37	l	54	2
4	E	21	V	38	m	55	3
5	F	22	W	39	n	56	4
6	G	23	X	40	o	57	5
7	H	24	Y	41	p	58	6
8	I	25	Z	42	q	59	7
9	J	26	a	43	r	60	8
10	K	27	b	44	s	61	9
11	L	28	c	45	t	62	+
12	M	29	d	46	u	63	/
13	N	30	e	47	v		
14	O	31	f	48	w	(pad)	=
15	P	32	g	49	x		
16	Q	33	h	50	y		

De Base64-tabel telt 64 (2^6) karakters en bestaat uit de karakters A-Z, a-z, 0-9, '+', '/'. Het '='-karakter wordt gebruikt voor opvulling wanneer het totaal aantal bits geen veelvoud is van 24 zodat er op het einde van de omzetting toch 4 tekstkarakters geretourneerd worden. De Base64-tabel is een subset van ASCII en heeft als voordeel dat alle karakters in alle ASCII en EBCDIC-varianten voorkomen. Dit geldt bijvoorbeeld niet voor BinHex, een alternatieve encodingswijze die vooral binnen Macintosh omgevingen wordt toegepast. Base64 gebruikt ook een andere tabel dan uuencode.

Binnen digitale archiveringsprojecten wordt Base64-encoding voornamelijk toegepast bij de XML-archivering van digitale archiefdocumenten. XML-elementen kunnen geen binaire tekens of gereserveerde karakters bevatten. Via Base64-encoding zorgt men ervoor dat de XML elementen enkel tekstkarakters bevatten. Op die manier kan een binair bestand in een XML-bestand worden ingekapseld. Het is aangewezen om expliciet te vermelden dat de inhoud van XML element een Base64 geëncodeerde bitstream bevat. Dit kan door Base64 te vermelden in de elementnaam of als attribuut (bijv. <bericht type="xs:base64Binary"> of <bericht encoding="base64">). De inhoud van Base64 geëncodeerd XML element is niet interpreteerbaar voor mensen en moet bij raadpleging eerst opnieuw gedecodeerd worden. Bij toekomstige decoding moet men minimaal over de werkwijze van het algoritme en de Base64-tabel beschikken. De Base64-tabel is de index tussen bits en overeenstemmende afdrubbare karakters. Hierdoor verliest het XML document wel een stuk van zijn autonomie.

Op het internet zijn diverse tools en broncodes beschikbaar voor Base64-omzettingen zodat Base64 vrij gemakkelijk in archiveringstools kan worden ingebouwd. Enige voorzichtigheid en uitvoerige tests zijn aangewezen zodat men er zeker van is dat de encoding/decoding exact verloopt zoals beschreven in RFC 2045.

Referenties: <http://www.ietf.org/rfc/rfc2045.txt>

5. BESTANDSFORMATEN

Voor de meeste computerapplicaties is een puur tekstuele opslag van de gegevens conform een bepaalde codetabel niet voldoende of efficiënt. Naast tekstkarakters worden ook binaire tekens opgeslagen. Veel softwareapplicaties gebruiken hun eigen methode om gegevens in een computerbestand te bewaren. Dit computerbestand heeft een vastgelegde structuur. Voor de correcte reconstructie van het document in het computerbestand gelden specifieke regels. Die structuur en regels worden het bestandsformaat genoemd.

Er bestaan verschillende soorten bestandsformaten. De bestandsformaten kunnen op diverse wijzen ingedeeld worden. ASCII-bestanden bevatten niets anders dan tekst- en controlekarakters. Binaire bestanden bevatten naast tekst- en controlekarakters ook binaire waarden. Men kan de

bestandsformaten ook indelen op basis van hun inhoud: tekst, afbeeldingen, audio, video, enz. Bepaalde bestandsformaten kunnen slechts één bepaald type gegevens of document bevatten. Andere bestandsformaten daarentegen kunnen verschillende soorten gegevens of documenten als inhoud hebben. Deze laatste categorie bestandsformaten worden doorgaans 'containers', 'wrappers', 'enveloppen' of 'metaformaten' genoemd.

Binaire formaten bestaan in het algemeen uit twee delen: een header en de data. De header bevat technische metadata over de data, maar biedt in veel gevallen ook de mogelijkheid om een aantal administratieve of archiefbeschrijvende metadata over het archiefdocument in het computerbestand te bewaren. Het data gedeelte van een binair formaat bevat de gegevens waaruit het archiefdocument wordt opgebouwd.

In onderstaand overzicht wordt van elk bestandsformaat de voornaamste eigenschappen 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 licenties aan te schaffen.

5.1 Tekstdocumenten

5.1.1 OFFICIËLE STANDAARDEN

5.1.1.1 Platte tekstbestanden (.txt)

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, tekstverwerkers, spreadsheetprogramma's, databanksoftware, webbrowsers, enz.). Bij het bewaren van teksten als platte tekstbestanden is het wel belangrijk om vast te leggen welke codetabel en eventueel welke versie werd gebruikt. Een ander belangrijk aandachtspunt zijn de controlekarakters voor de geregeleinden in het ASCII-bestand. Een geregeleinde kan immers met een carriage return (ASCII-waarde 13), een linefeed (ASCII-waarde 10) of beiden worden aangegeven. Dit is doorgaans platformafhankelijk. Windowssystemen gebruiken doorgaans de combinatie van een 'carriage return' en 'linefeed' om het einde van een regel aan te duiden, UNIX de 'linefeed' en Apple de 'carriage return'. Bij Unicode is dit overbodig.

In een plat tekstbestand wordt in de eerste plaats de inhoud van een tekstueel archiefdocument bewaard. In een plat tekstbestand is het doorgaans ook mogelijk om de structuur van een document te bewaren. Daartoe worden ASCII-karakters als veldscheidingstekens gebruikt. In de meeste gevallen wordt een tab-teken (ASCII-waarde 9) of een punt-komma (ASCII-waarde 59) als delimiter 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. Een geregeleinde wordt als scheidingstekens 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 tekst- en enkele controlekarakters bewaard. Alle binaire tekens gaan verloren.

Een plat tekstbestand is software-onafhankelijk. In die zin beantwoordt het volledig aan de archivalische noden. Platte tekstbestanden worden dan ook gebruikt voor de archivering van

eenvoudige tekstdocumenten en tabellen uit databanken. 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.

5.1.1.2 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 delimiters 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 schuine streep eraan toegevoegd op de tweede positie. 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>15 oktober 2005</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 worden gebruikt.

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 toe te voegen werd in 1992 een nieuwe standaard door ISO vastgelegd: HyTime (ISO/IEC-10744(1992): *Information technology -- Hypermedia/Time-based Structuring Language*).

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

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 uitgeverwereld 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

omzettingmoment 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.

5.1.1.3 HyperText Markup Language (.htm, .html): 4.01

Hypertext Markup Language (HTML) is de markuptaal waarin pagina's op het World Wide Web worden 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, opsommingen, tabellen, enz. Aan de tags worden attributen toegekend. Webpagina's worden geraadpleegd met een webbrowser. HTML-bestanden zijn platte tekstbestanden en kunnen dus met een gewone teksteditor of tekstverwerker worden geëditeerd.

HTML werd vastgelegd door het *World Wide Web Consortium (W3C)* en heeft algemeen de status van defacto standaard. Dit is ondermeer het geval door de wijdverspreide HTML-versies 2.0 (IETF, 1994), 3.2 (1996), 4.0 (1997), en 4.01 (1999). Deze laatste versie is overgenomen door ISO en werd als officiële standaard vastgelegd door (ISO-15445 (2000): *Information Technology. Document description and processing languages. HyperText Markup Language*). Andere HTML-versies dan deze 4.01-versie zijn dus geen officiële, maar defacto standaarden. De recentste HTML-versie is 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 de HTML-compatibiliteit regels worden toegepast. Net zoals bij HTML zijn er drie varianten van 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 niet door de gebruiker kan worden uitgebreid. Bij elke HTML versieverhoging wordt het assortiment tags en attributen 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. Dit is bijvoorbeeld het geval met de dynamic HTML-uitbreidingen die Netscape (dhtml) en Microsoft (DHTML) op de HTML-standaard maakten. Beide uitbreidingen zijn niet zonder meer uitwisselbaar en leveren leesbaarheidsproblemen op. Het W3C stelt tools ter beschikking voor het valideren van webpagina's of het 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 populariteit leidde al gauw tot oneigenlijke toepassingen van HTML. Met het oog op toevoegen van lay-out werden nieuwe tags ontworpen (,
, 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)⁵. Net zoals de afbeeldingen zijn stylesheets meestal afzonderlijke computerbestanden, al kunnen ze ook in de HTML-pagina worden ingebed.

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. Een aantal initiatieven voor websitesarchivering voegen HTML-pagina's in ARC-bestanden (zie 5.1.2.9). HTML kan ook gebruikt worden bij de 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/>

5.1.1.4 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 staat voor Open Document Architecture, maar af en toe wordt ODA ook gedefinieerd als Office Document Architecture. 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 softwareproducten en de concurrentie van SGML⁶. 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>

5.1.2 DEFACTO STANDAARDEN

5.1.2.1 eXtensible Markup Language (.xml)

XML werd in 1998 als een Recommendation van het *World Wide Web consortium* gepubliceerd. XML is eigenlijk geen bestandsformaat, maar een taal waarvan de grammatica en spellingsregels in de XML-Recommendation is vastgelegd. Inmiddels is in februari 2004 versie 1.1 gepubliceerd.

De XML-taal is grotendeels gebaseerd op SGML, maar is niet zo complex of uitgebreid. De syntaxregels van XML zijn heel eenvoudig. Net zoals SGML is XML een markuptaal, wat inhoudt dat

⁵ Voor meer informatie over CSS, zie: <http://www.w3.org/Style/CSS/>

⁶ <http://www.infoma.jyu.fi/digimedi/Pasi/eds/odaodif.htm>. ODA stond vroeger voor Office Document Architecture.

documenten met tags worden gemarkeerd. In XML hebben de tags een betekenisgevende functie, terwijl in HTML bijvoorbeeld de tags een presentatiedoel hebben. In de HTML-standaard liggen de tags en hun attributen vast. Bij XML kan de gebruiker zelf zijn tags en attributen samenstellen. XML is immers uitbreidbaar. Door semantische tags te gebruiken, kan men de documentcomponenten identificeren. De XML-tags worden ook gebruikt als veldscheidingstekens en voor het structureren van een document.

Door de semantische tags te structureren of te nesten, kan men op een zeer doorgedreven wijze XML-documenten op een expliciete en menselijk leesbare wijze modelleren. Dit is het basisprincipe van XML. Hierdoor kan men niet alleen kennis in tijd overbrengen, maar ook de interne structuur en de semantiek van een document op een expliciete wijze archiveren. Er wordt dan ook van XML-documenten gezegd dat ze autonoom of zelfvoorzienig zijn.

De XML-specificatie maakt een onderscheid tussen twee soorten XML-documenten: *well-formed* (welgevormd) en *valid* (geldig). Welgevormde documenten zijn conform de XML-specificatie samengesteld en passen de XML-regels correct toe. Elk XML-document moet welgevormd zijn. De geldige XML-documenten zijn niet alleen welgevormd, maar zijn ook conform een specifieke documentstructuur samengesteld. Die documentstructuur wordt in een DTD (Document Type Definition) of een XML Schema gedefinieerd. XML erfde het DTD-mechanisme over van SGML, terwijl een XML Schema zelf een XML-document is. Met een XML Schema kan de documentstructuur gedetailleerder worden gedefinieerd dan met een DTD (bijv. aantal keren dat een element voorkomt, datatypen van de elementen, enz.).

XML-documenten bevatten geen opmaak. De opmaak wordt in afzonderlijke CSS- of XSL-stylesheets opgeslagen. Hierdoor worden inhoud en structuur enerzijds en opmaak anderzijds van elkaar gescheiden. XML-documenten bevatten evenmin enige businesslogica. Dit alles samen zorgt ervoor dat XML-documenten volledig platformafhankelijk zijn. Voor het verwerken van XML-documenten zijn twee technologieën beschikbaar: DOM en SAX.

In essentie zijn XML-documenten niets meer dan platte tekstbestanden. XML-documenten kunnen bijgevolg alleen tekstkarakters bevatten. Een aantal tekstkarakters zijn bovendien gereserveerd omdat ze een welbepaalde functie vervullen. Toch is het mogelijk om binaire waarden in XML-documenten op te nemen. Men kan binaire waarden via Base64 naar tekstkarakters omzetten of ze als als CDATA-sectie definiëren. XML-documenten zijn editeerbaar en raadpleegbaar met elk computerprogramma dat in staat is om een ASCII-bestand te openen.

Rond XML is een hele XML gerelateerde technologiefamilie ontstaan. Het *World Wide Web consortium* heeft naast XML nog diverse andere aanbevelingen geformuleerd waarmee XML kan gecombineerd worden: XSL(T) (transformeren en publiceren van XML-documenten), XLink (linken van XML-documenten), XQuery (bevragingstaal), XPointer (identificeren van URI-verwijzingen naar informatiebronnen op het web), XPath (adresseren van onderdelen van XML-documenten), enz. Deze gerelateerde technologieën liggen mee aan de basis van de populariteit van XML.

XML is op diverse wijzen inzetbaar bij de archivering van digitale archiefdocumenten. De archiveringscasussen waarbij XML wordt gebruikt zijn dan ook legio. XML is bruikbaar als uitwisselings-, metadata-, archiverings- en inkapselingsformaat.

Als uitwisselingsformaat is XML uitermate geschikt voor de uitwisseling van archiefdocumenten en hun metadata tussen archiefvormer – archief – archiefgebruiker, ook al werken deze partijen met niet compatibele informatiesystemen.

De metadata over archiefbestanddelen kunnen als XML-documenten worden vastgelegd. Dit biedt het voordeel dat de metadata op een expliciete en platformafhankelijke wijze worden vastgelegd en achteraf gemakkelijk verwerkt kunnen worden. Men kan volledig in functie van de eigen noden en behoeften een eigen metadataschema in XML toepassen.

XML kan ten derde ook gebruikt worden als archiveringsformaat voor digitale archiefdocumenten. XML beantwoordt immers aan nagenoeg alle criteria van een geschikt archiveringsformaat. XML wordt

internationaal als geschikt archiveringsformaat voor (semi-) gestructureerde tekstuele documenten zoals e-mail en databanken naar voor geschoven. XML is ook bruikbaar voor de archivering van tekstverwerkingsdocumenten. Bijzondere aandacht hierbij dient wel uit te gaan naar het bewaren van de originele opmaak. De originele opmaak kan slechts in XML worden overgenomen als hier van bij de creatie al rekening werd gehouden.

Ten slotte is XML ook bruikbaar als inkapselingsformaat. In één en hetzelfde XML-document kan men zowel de metadata als het eigenlijke archiefdocument bewaren. Dit heeft als voordeel dat het archiefdocument en zijn metadata onlosmakelijk met elkaar verbonden zijn. Bovendien is XML uitbreidbaar en is men voor de metadata niet gebonden aan een voorgedefinieerd metadataset in de fileheader. Het archiefdocument kan in XML en/of in zijn oorspronkelijk bestandsformaat in de XML-container worden opgenomen. Zo kan één XML-container de metadata en de verschillende verschijningsvormen van hetzelfde archiefdocument bevatten (zie bijv. de opslagmethode voor digitale archiefdocumenten die eDAVID ontwikkelde en die door de stad Antwerpen wordt toegepast). Een andere toepassingsmogelijkheid van XML als inkapselingsformaat is de opname van alle gerelateerde documenten in één en hetzelfde XML-bestand (zie bijv. de *Persistent Object Preservation*-methode van het NARA).

Referentie: <http://www.w3c.org/>; <http://www.oasis-open.org/>; <http://xml.coverpages.org/>; F. BOUDREZ, <XML/> en digitaal archiveren, Antwerpen, 2002. (http://www.edavid/davidproject/teksten/XML_digitaalarchiveren.pdf); F. BOUDREZ, *Digitale containers voor het digitale archiefdepot*, Antwerpen, 2005 (http://www.edavid.be/docs/digitale_containers.pdf)

5.1.2.2 Tagged Image File Format (.tif, .tiff)

Tekstdocumenten kunnen perfect als TIFF-documenten worden gearhiveerd. Tekstdocumenten worden hierbij als een afbeelding opgeslagen en kunnen bijgevolg niet meer geëditeerd worden. Voor de omzetting van een tekstdocument naar een TIFF-bestand dient men over een TIFF image driver te beschikken (vergelijkbaar met een PDF-driver). MS Office 2003 is standaard uitgerust met een zo'n image driver, maar past altijd compressie toe.

Voor het omzetten van tekstdocumenten die meerdere pagina's beslagen, heeft men meerdere opties. Het volledige document kan als één multi-page TIFF-bestand worden bewaard. Een andere mogelijkheid is dat elke pagina als een single-page TIFF wordt opgeslagen en dat alle TIFF-afbeeldingen die één document vormen in een XML-bestand worden verpakt.

TIFF wordt in verschillende landen als archiveringsformaat voor tekstuele documenten gebruikt. In Zwitserland bijvoorbeeld worden alle officedocumenten als TIFF-documenten overgedragen naar het Bundesarchiv. TIFF is ook het voorgeschreven bestandsformaat waarin tekst en afbeeldingen naar het Deens Nationaal Archief worden overgedragen. In Denemarken hanteert men ook de vereiste dat alle archiefdocumenten binnen informatiesystemen als TIFF-documenten archiveerbaar moeten zijn. Voor meer info over TIFF: zie 5.2.1.1.1.

5.1.2.3 HyperText Markup Language (.htm, .html)

De niet officieel gestandaardiseerde versies van HTML worden als defacto standaarden beschouwd. Dit geldt voor zowel de W3C HTML-standaard als enkele producenteigen HTML-uitbreidingen (Microsoft, Netscape, enz.). Het spreekt voor zich dat de voorkeur uitgaat naar de ISO en W3C gestandaardiseerde HTML-versies en dat producenteigen implementaties met hun softwareafhankelijke uitbreidingen worden vermeden. Voor meer info: zie 5.1.1.2 HyperText Markup Language.

5.1.2.4 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 platformafhankelijke 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 procedurdefinities) 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 compressieloos worden toegepast. Een PostScriptbestand kan zowel een statisch of een dynamisch formaat zijn.

PostScript is bruikbaar voor de uitwisseling of opslag van afdrubbare 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.

PostScript wordt niet meer veel gebruikt voor de digitale archivering van archiefdocumenten. Het courante archiefgebruik dateert ondertussen al van een tiental jaren geleden. De archiefbeherende instellingen die PostScript als archiveringsformaat toepassen, zijn inmiddels naar PDF overgeschakeld.

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

5.1.2.5 Portable Document Format (.pdf)

Softwareproducent Adobe Systems startte de verspreiding van PDF midden 1993. PDF is eigendom van Adobe. Adobe System is meer bepaald drager van het auteursrecht op PDF. Alhoewel Adobe daartoe niet verplicht is, publiceert ze de technische specificatie van het PDF-formaat op haar website zodat andere ontwikkelaars zelf PDF-toepassingen kunnen programmeren voor het bekijken of produceren van PDF-documenten.

Er bestaan verschillende versies van de PDF-specificatie. De PDF-versies mag men niet verwarren met de verschillende versies van de Adobe Acrobatsoftware die wordt gebruikt om PDF-documenten samen te stellen. Acrobat versie 5.0 en Illustrator 9.0 zijn op specificatie 1.4 gebaseerd. Acrobat 6.0 ondersteunt PDF-versie 1.5 en Acrobat 7.0 creëert PDF-bestanden conform PDF versie 1.6. De recentste Acrobatversies kunnen even goed PDF-documenten op basis van een lagere PDF-versie maken. Het PDF-versienummer wordt als commentaar aangegeven in het begin van een PDF-bestand (bijv. “%PDF-1.4”) of men kan dit opvragen via de documenteigenschappen. De voornaamste kenmerken van elke PDF-versie zijn:

- versie 1.1 (1995): encryptie, artikels, omzettingen, verbeterde hyperlinks, acties
- versie 1.2 (1997): prepress functionaliteiten, externe streams, flate compressie, formuleren, Aziatische fonts, meer annotaties
- versie 1.3 (1999): PostScript 3 imaging model, ICC color, logische structuur, JavaScript, meer annotaties, digitale handtekeningen
- versie 1.4 (2001): transparantie, 128-bit encryptie, XML metadata, Tagged PDF (toegankelijkheid)
- versie 1.5 (2003): JPEG2000-compressie, Object Stream compressie, PDF/X
- versie 1.6 (2004): ondersteuning van de AES-encryptiestandaard, uitbreiden van de annotatiemogelijkheden, inbedding van bijlagen, opname van 3D-modellen, verbetering van de PDF-tagging

De basiseenheid van een PDF-bestand is een blanco blad. Adobe wil immers met PDF een brug bouwen tussen de papieren en digitale wereld. PDF is gebaseerd op PostScript (*.ps-bestanden, zie 5.1.2.4)⁷. De data in het PDF-document bepalen op welke plaatsen er “inkt” van om het even welke kleur komt, welke marges er zijn, welke fontspecificaties worden gebruikt, enz. De manier waarop de informatie in PDF-bestanden wordt opgeslagen, wordt het Adobe imaging model genoemd. 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 platformafhankelijk bestandsformaat voor. Hiermee bedoelt men eigenlijk dat hetzelfde PDF-bestand op verschillende besturingssystemen raadpleegbaar is. Voor het openen van PDF-bestanden moet men wel over een viewer beschikken. Men kan hiervoor de Acrobat-Reader van Adobe of een (freeware) viewer van een andere producent gebruiken. GhostScript en GhostView zijn voorbeelden van open source computerprogramma's voor het maken van PDF-bestanden⁸. Er zijn ook een aantal websites die het maken van PDF-bestanden als on line webservice aanbieden.

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-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.

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

⁷ 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. Postscriptbestanden bevatten alle informatie van de documentpagina's, informatie over de gekoppelde bestanden (bijv. geïmporteerde illustraties), lettertypen en printerinstructies. Postscript is een geschikt bestandsformaat om bijvoorbeeld PageMakerbestanden op een platformafhankelijke wijze te archiveren.

⁸ <http://www.cs.wisc.edu/~ghost/>

tekst in een PDF-bestand wordt niet als ASCII- of Unicodekarakters opgeslagen⁹. 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.

PDF-documenten kunnen ook meer bevatten dan tekst, grafieken en afbeeldingen. Ook audio, video en kleine computerprogramma's kunnen in PDF-documenten worden ingebed. Vanuit archiveringsperspectief is dit echter af te raden.

Men kan PDF-documenten maken met het Acrobatprogramma van Adobe of met software van andere producenten. 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 Acrobat PDF-writer maakt deel uit van Acrobat 4 (standaard) en 5 (optioneel). Acrobat 6 en 7 zijn niet meer met een PDF-writer uitgerust. 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). Wanneer een tekstdocument enkel tekst bevat, dan is de bestandsomvang van een PDF-bestand doorgaans kleiner dan van een Word- of Postscript-bestand. Tekstdocumenten met ingevoegde afbeeldingen zijn na omzetting naar PDF dikwijls groter.

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 herkennen en correct kunnen weergeven. Bij (on)gestructureerde PDF-bestanden is dit niet of veel minder het geval. Getagde PDF-documenten bieden de beste omzettingresultaten naar andere formaten of apparaten (bijv. e-Bookreader) en worden door screenreaders het betrouwbaarst weergegeven. 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 drie PDF-specificaties voorzien in de mogelijkheid om bestanden te taggen of te structureren. Deze functionaliteit werd echter pas met Acrobat 5.0 geïntroduceerd. Met een speciale plug-in kan men ongestructuurde PDF-documenten omzetten naar getagde PDF-documenten. Acrobat 6 en 7 biedt standaard deze functionaliteit. Het taggen van PDF-documenten kan onmiddellijk bij de omzetting of achteraf gebeuren.

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 bood enkel de mogelijkheid om PDF-documenten als PostScriptbestanden te exporteren. Versie 5 laat de gebruiker toe het PDF-bestand als RTF-bestand te bewaren zodat het in een tekstverwerkingsprogramma verder kan worden bewerkt. Acrobat 6 en 7 laten toe dat de PDF-

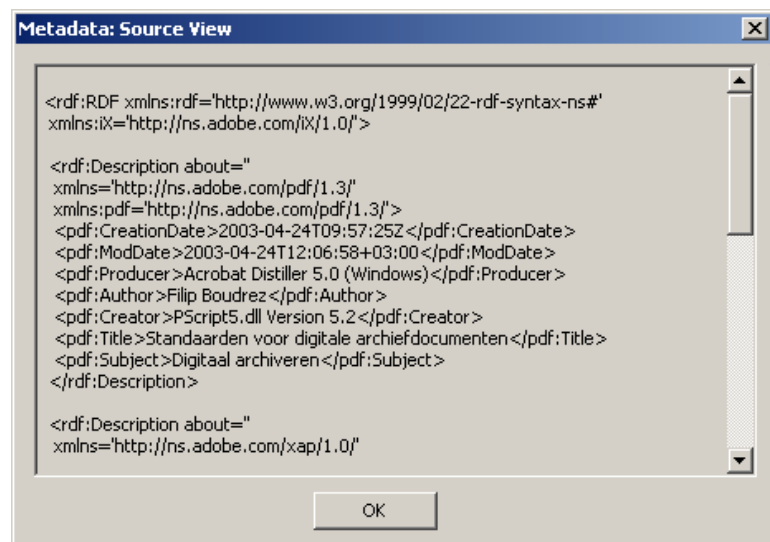
⁹ De optie 'ASCII-formaat' in de conversiesettings kan tot verwarring leiden. Hiermee wordt niet bedoeld dat de tekst als ASCII-karakters worden vastgelegd. ASCII-formaat wordt hier gebruikt als tegenovergestelde van binair bestand. Deze optie wordt het best gebruikt bij uitwisseling of om de bewerking van een PDF-bestand in een teksteditor mogelijk te maken.

bestanden rechtstreeks kunnen geëditteerd worden. Wie een beetje handig is met een grafisch programma zoals Photoshop kan ook PDF-bestanden manipuleren. PDF-bestanden kunnen ook relatief gemakkelijk worden omgezet naar ASCII, RTF, XML of HTML. Dit kan in Acrobat 6, 7 of met aparte tools¹⁰.

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.

PDF en Acrobat bieden een aantal XML-functionaliteiten. PDF gebruikt XML ten eerste voor de opslag van documentmetadata. Dit mechanisme wordt XMP of voluit XML Metadata Platform genoemd. XMP is een soort combinatie van de XML- en RDF-standaarden. Het documentprofiel wordt binnen het PDF-bestand in XML bijgehouden. Deze XML-metadata kunnen met om het even welke teksteditor worden gewijzigd. Getagde PDF-documenten kunnen ten tweede als XML-documenten worden geëxporteerd. Acrobat 5 heeft hiervoor nog wel een plug-in nodig, maar Acrobat versie 6 en 7 voorzien standaard deze functionaliteit.

Afbeelding 2: XML documentprofiel van het PDF-bestand "standaarden.pdf".



De meeste gegevens die nodig zijn om documenten in hun originele verschijningsvorm af te drukken of te presenteren, kunnen in het PDF-bestand zelf bewaard. Hierdoor wint het document aan autonomie en kan het PDF-bestand afgedrukt of op het scherm gepresenteerd 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. Niet elk font kan echter zomaar in het PDF-document worden ingebed. 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.

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 resoluatieloos 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 voorgeschreven (art.

¹⁰ <http://www.ra.informatik.uni-stuttgart.de/~gosho/pdfhtml/index.html>; <http://www.mosarca.com/bva-myfra/trap2gb.htm>; <http://www.cs.wisc.edu/~ghost/>

6). De VERS-archiveringsstrategie voor digitale archiefdocumenten is in grote mate op PDF gebaseerd. Ook het NARA accepteert de overbrenging van documenten die permanent worden bewaard in PDF.

Toch zijn er een aantal belangrijke kanttekeningen. PDF is ten eerste niet bruikbaar voor alle types tekstuele documenten. PDF is hoofdzakelijk 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. Archiefdocumenten zonder een papieren equivalent zoals een database kunnen niet zomaar als PDF-document worden gearhiveerd. PDF heeft ten tweede ook als kenmerk dat het heel flexibel en op uiteenlopende wijzen bruikbaar is. Dit wordt vanuit commercieel standpunt als een groot voordeel voorgesteld, maar kan voor archivering ernstige problemen opleveren. Niet elke toepassingsmogelijkheid binnen PDF is immers aangewezen voor bewaring op lange termijn. (bijv. gebruik van gepatenteerde compressies, kleurenschema's of fonts, encryptie, beveiliging, afhankelijkheid van externe bronnen, enz.). Met andere woorden, onbeperkte PDF-documenten zijn niet geschikt voor archivering. PDF-documenten kunnen dus wel voor lange termijnarchivering in aanmerking komen, maar alleen als de bestanden met goede instellingen vanuit archiveringsstandpunt werden gecreëerd. PDF-bestanden hebben ten derde ook geen volwaardige ASCII-basis waardoor de afhankelijkheid van aangepaste software relatief groot is. Een vierde nadeel is een gevolg van de evolutie van de PDF-specificatie. De specificatie van het bestandsformaat wordt aan een snel temp gewijzigd (drie versies op vier jaar tijd), en kan dus bezwaarlijk als stabiel gekenmerkt worden. Bovendien is de specificatie vrij complex (1236 pagina's!). In theorie is het ten allen tijde mogelijk om op basis van de gepubliceerde technische specificatie een nieuwe PDF-viewer te ontwikkelen, maar in de praktijk kan dit een serieuze uitdaging inhouden. Vooral de ingewikkelde PDF-tags doen nogal wat problemen rijzen.

Het belangrijkste nadeel van PDF 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. De algemene tendens is momenteel om PDF als een geschikt formaat voor archivering op middellange termijn te beschouwen. Voor lange termijnarchivering is PDF wellicht niet veilig genoeg.

Mede vanwege het probleem van de producentgebondenheid wordt momenteel werk gemaakt van de officiële standaardisatie van PDF. Het standaardiseringsinitiatief gaat uit van AIIM en wordt sterk gesteund door de Amerikaanse overheid. Het initiatief heeft niet tot doel om het hele PDF-formaat als standaard vast te leggen, maar een subset te definiëren die geschikt is voor archiveringsdoeleinden. Deze subset kan enigszins vergeleken worden met een profiel dat is gebaseerd op de instellingen die aangewezen zijn vanuit archiveringsstandpunt. Een andere officieel gestandaardiseerde subset van PDF, nl. PDF/X (PDF for eXchange), dient hierbij als voorbeeld (ISO 15930-1(2001): *Graphic technology -- Prepress digital data exchange -- Use of PDF -- Part 1: Complete exchange using CMYK data (PDF/X-1 and PDF/X-1a)* en ISO 15930-3(2002): *Graphic technology -- Prepress digital data exchange -- Use of PDF -- Part 3: Complete exchange suitable for colour-managed workflows (PDF/X-3)*). PDF/X is een subset bedoeld voor prepress bestandsuitwisseling. De subset voor archiveringsdoeleinden wordt gebaseerd op versie 1.4 van het PDF-bestandsformaat en kreeg de voorlopige naam PDF/A (PDF for Archiving) mee.

In de subset worden een aantal instellingen en eigenschappen van het PDF-formaat vastgelegd zodat een getrouwe reconstructie van het archiefdocument op lange termijn mogelijk zal zijn. De bedoeling is PDF/A-documenten te creëren die later gemakkelijk kunnen gemigreerd worden of waarvoor gemakkelijk viewers kunnen worden ontwikkeld. PDF/A zal zo weinig mogelijk afhankelijk zijn van externe bronnen en zal een aantal beperkingen inhouden. Kenmerken van PDF/A zijn:

- alle lettertypes worden ingesloten
- eigendomsgebonden fonts of compressietoepassingen (bijv. LZW en JBIG 2) zijn niet toegelaten
- encryptie, paswoorden en andere vormen van beveiliging zijn niet toegelaten
- geen inbedding van audio, video, (Java)scripts of kleine executables
- inbedding van metadata op basis van XMP
- automatisch bijhouden van een audit trail binnen het document (bijv. wie heeft het document geopend, welke annotaties, enz.).

- beperking op de kleurschema's, toepassen van gestandaardiseerde kleurschema's
- geen afbeeldingen met transparantie

Er wordt ook gedacht aan een vorm van validatie van PDF-documenten die voldoen aan de PDF/A-standaard. Net zoals bij PDF/X zal in de standaard niet alleen de samenstelling van PDF/A worden vastgelegd, maar ook de eigenschappen en het gedrag van readers.

Met de standaardisatie van PDF/A wijzigt ook de status van PDF voor archiveringsdoeleinden. PDF/A is dan niet langer meer in handen van één producent, maar wordt beheerd door een groep die controle houdt op de samenstelling en mee de lange termijn leesbaarheid verzekert. Dit biedt in ieder geval een grote zekerheid inzake digitale duurzaamheid dan een louter gepubliceerde specificatie van een bestandsformaat. Wel valt nog af te wachten of (onderdelen van) PDF/A volledig patentvrij zijn. De ISO-norm vermeldt immers heel uitdrukkelijk dat "some elements of this document may be the subject of patent rights".

Producent Microsoft kondigde in augustus 2005 aan de nieuwe versie van zijn besturingssysteem ('Vista') en kantoor suite ('Office 12') standaard uit te rusten met een viewer en een writer voor 'Metro' of het XPS ('XML Paper Specification') Document Format. Dit formaat moet een alternatief voor PDF worden.

Referentie: <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>; <http://www.aiim.org/standards.asp?ID=25013>; *PDF as a standard for archiving. White paper Adobe.*

5.1.2.6 OpenOffice XML - OpenDocument

Het open source officepakket OpenOffice gebruikt een op XML gebaseerd bestandsformaat voor de opslag van alle documenten (tekstverwerking, spreadsheets, presentaties, enz.) die met het dit pakket worden gecreëerd. Dit bestandsformaat is ontwikkeld door het open source initiatief 'OpenOffice.org' en de bestandformaat specificatie is open en gedocumenteerd. Op basis van het OpenOffice-bestandsformaat heeft de standaardiseringsorganisatie OASIS een nieuw standaardformaat voor kantoor documenten vastgelegd (mei 2005). Deze standaard kreeg de naam OpenDocument mee. Inmiddels is dit formaat ook voor standaardisatie bij ISO ingediend. De algemene verwachting is dat OpenDocument vrij snel een officiële ISO-standaard wordt. 'OpenOffice.org' engageerde zich alvast om OpenDocument als basisbestandsformaat voor OpenOffice 2.0 te gebruiken. Ook KOffice is van plan om OpenDocument als bestandsformaat aan te nemen. Microsoft is (voorlopig?) niet van plan om het OpenDocument-formaat als applicatieformaat voor zijn Officepakket (versie 12) te hanteren.

Het OpenOffice-bestandsformaat is gebaseerd op XML. In feite is het OpenOffice-bestandsformaat een gecomprimeerd bestand dat onder andere een verzameling XML-bestanden met ingebedde stylesheets bevat. Een tekstbestand bevat bijvoorbeeld naast een bestand met de mime-typing: content.xml, meta.xml, settings.xml, styles.xml en een manifest.xml. Voor de compressie wordt ZIP gebruikt. Afbeeldingen worden als afzonderlijke bestanden bewaard en mee in het SXW-formaat verpakt.

Het archiveren van archiefdocumenten als OpenOffice-bestanden biedt het voordeel dat gearchiveerde officedocumenten volledig functioneel blijven, al is dat geen absolute vereiste vanuit archiveringsperspectief. Een ander interessant punt is dat één archiveringsformaat voor verschillende documenttypen kan worden gebruikt. Nadelen van het gebruik van XML als archiveringsformaat zijn de complexiteit van de XML-structuur, de compressie en de afhankelijkheid van OpenOffice.org. Dit laatste houdt voorlopig niet alleen de afhankelijkheid aan het OpenOffice-pakket in, maar ook de afhankelijkheid van het open source initiatief. De kans is wel vrij groot dat het OpenOffice-formaat uitgroeit tot een officiële standaard, maar wanneer en in welke vorm dit zal gebeuren, is nog niet duidelijk. De kans is echter klein dat OpenDocument volledig identiek aan SXW zal zijn. Voorlopig is OpenOffice dus een standaard-in-wording. Het lijkt aangewezen om te wachten tot OpenDocument

officieel als standaard wordt vastgelegd, alvorens het als lange termijn archiveringsformaat te gaan gebruiken.

Het gebruik van het OpenOffice-formaat als archiveringsformaat is niet wijdverspreid. De National Archives van Australië zijn tot op heden de enige archiefinstelling die SXW volop als archiveringsformaat gebruiken. Zij ontwikkelden de migratietool XENA voor de migratie van bijv. MS Office-documenten naar het OpenOffice-formaat.

Referentie: <http://www.openoffice.org>; <http://xml.openoffice.org/>

5.1.2.7 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 laatste gedocumenteerde 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 gegevensuitwisseling 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/dnrftspec/html/rftspec.asp>.

5.1.2.8 MS Word (.doc)

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 de verspreiding op grote schaal van het tekstverwerkingsprogramma MS Word. De recentste versies zijn Word 6.0, Word 97 en Word 2000, 2002 en 2003. De applicatie MS Word is beschikbaar voor de verschillende courante platformen. Uitwisseling van hetzelfde Wordbestand tussen verschillende besturingssystemen is echter 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 bitstreams 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 Wordversies kunnen bestanden inlezen die werden aangemaakt in Word 6.0 en Word 97. Het correct inlezen van nog oudere versies wordt al wat moeilijker. MS Worddocumenten zijn versiegebonden. De gebruiker stelt dit niet altijd vast, maar elke MS Wordversie hanteert zijn eigen versie van het bestandsformaat. Hierdoor kan het openen van een Word 2003-document met MS Word 2000 problemen opleveren.

Vanaf versie 2003 (MS Office nr. 11) gebruikt Microsoft een op XML gebaseerd bestandsformaat. De XML Schemas van dit documentformaat worden vrijgegeven, maar volgens open source ontwikkelaars zijn deze schemas niet zo "vrij" en "open" als ze zouden moeten zijn. Microsoft heeft immers patenten op het gebruik van XML-bestandsformaten in kantoorapplicaties aangevraagd en ontwikkelaars die het formaat in hun eigen applicaties willen gebruiken, dienen een licentieovereenkomst te tekenen waarin bepaald gebruik wordt verboden. Bovendien bevat dit bestandsformaat nog steeds binaire gegevens, wat de vrije uitwisseling in de weg staat.

Wordbestanden zijn niet geschikt voor de bewaring op lange termijn. De ondersteuning is beperkt in tijd en volledig in handen van één producent. Bovendien hanteert producent Microsoft sinds oktober 2002 een "life cycle support"-politiek. Softwarepakketten worden maar een beperkte tijd meer ondersteund. Zo eindigt de "mainstream support" voor Word 2003 al op 31 december 2008. 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.

Microsoft Corporation verspreidt eveneens een viewer voor Word 97 en 2003-bestanden.

Referentie: /

5.1.2.9 ARC (.arc)

ARC is het bestandsformaat waarin verschillende initiatieven voor websitesarchivering webpagina's in archiveren. ARC werd ontwikkeld door het bedrijf Alexa en de non-profit organisatie 'The Internet Archive' zodat miljoenen webpagina's op een efficiënte wijze binnen een regulier bestandssysteem kunnen beheerd worden. ARC-bestanden zijn een aggregatie van webpagina's en zijn doorgaans 100 MB groot. Men mag ARC niet verwarren met het oude compressie- en verpakkingsformaat 'Archive' (LH Arc, SQUASH).

Een ARC-bestand bestaat uit een verzameling webpagina's. Elke webpagina bestaat uit twee blokken: de header en de data. De header bevat de URI van het document, het IP-adres, de timestamp van het archiveringstijdstip, de mime-typing van de inhoud, de result-code of response-code, een checksum, de plaatsaanduiding binnen het ARC-bestand, de bestandsnaam van het ARC-bestand en de omvang van het datablok. Deze metadata worden automatisch geëxtraheerd uit de HTTP-header en aangevuld met automatisch gegenereerde metadata op het tijdstip van archivering. De data van de webpagina volgen op de header en worden gemarkeerd met HTML-tags.

De ARC-specificatie is open en gedocumenteerd. ARC is een heel eenvoudig op ASCII gebaseerd bestandsformaat dat gemakkelijk aan lokale behoeften kan aangepast worden. Zo breidde het Deense netarchive.dk-project de metadatavelden van het ARC-formaat uit. Inmiddels werd ook versie 1.0 van het ARC-formaat voorzien van een XML-schema voor metadata (<http://archive.org/arc/1.0/xsd>).

ARC-bestanden kunnen samengesteld worden door de Heritrix- of de HTTrack-webcrawler. Deze laatste webharvester heeft hiervoor wel een plug-in nodig. Doorgaans wordt gzip-compressie toegepast bij het schrijven van ARC-bestanden, al kan dit uitgeschakeld worden. Het NedLib-project ontwikkelde een tool die de bestanden vastgelegd door de NedLib-harvester naar ARC omzet. Ook voor het raadplegen van ARC-bestanden is bijzondere software nodig. Momenteel zijn hiervoor verschillende tools beschikbaar: ARCReader, libarc, ProxyViewer van netarchive.dk of de ARC Reader van de Franse Nationale Bibliotheek.

In theorie zijn ARC-bestanden volledig zelfvoorzienig en niet afhankelijk van externe bronnen. In de praktijk echter is het gebruik van een externe databank een meerwaarde om snel webpagina's binnen ARC-bestanden terug te vinden.

Referentie: M. BURNER en B. KAHLE, *Arc file format*, september 1996 (<http://www.archive.org/web/researcher/ArcFileFormat.php> en <http://pages.alexacompany.com/company/arcformat.html>); S.S. CHRISTENSEN, *Archival Data Format Requirements*, 2004. (http://www.netarchive.dk/website/publications/Archival_format_requirements-2004.pdf)

5.2 Afbeeldingen

De bestandsformaten voor de opslag van digitale afbeeldingen worden doorgaans in drie groepen verdeeld: de bitmap/rasterformaten, de vectorformaten en de meta-formaten.

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, de gehanteerde kleurschema's, hun mogelijke ondersteuning van transparantie en interlacing, 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 één pixel en bijgevolg de afbeeldingen kan bevatten (1-bit kleurdiepte: zwart/wit (monochroom); 4-bit kleurdiepte: 16 kleuren (grijswaarden); 8-bit kleurdiepte: 256 kleuren (kleur); 24-bit kleurdiepte: 16.777.216 kleuren (ware kleuren)). De kleuren worden geïdentificeerd door een numerieke waarde die verwijst naar een bepaald kleurschema. De meest voorkomende kleurschema's zijn RGB (schermweergave) en CMYK (afdrukken). Sommige formaten ondersteunen transparantie en interlacing. Transparantie houdt in dat (bepaalde delen van) de afbeelding doorschijnend zijn. Hiervoor wordt het zogenaamde alfakanaal gebruikt. Interlacing laat toe dat de afbeelding eerst grof wordt weergegeven terwijl de resterende bits of tussenliggende rasterlijnen worden ingeladen. 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.

De meta-formaten kunnen in één computerbestand zowel een raster- als vectorversie van dezelfde afbeelding opslaan. De meta-formaten hebben als doel uitwisseling tussen applicaties en besturingssystemen gemakkelijk te laten verlopen.

5.2.1 AFBEELDINGEN IN META-FORMAAT

5.2.1.1 Officiële standaard

5.2.1.1.1 *Computer Graphics Metafile (.cgm)*

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 raster en vectorafbeeldingen. 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 platformafhankelijk.

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 systeemafhankelijke 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 automobielenindustrie. Standaard CGM wordt compressieloos toegepast.

Referentie: <http://www.cgmpopen.org>; <http://www.iso.ch>; <http://www.w3.org/TR/REC-WebCGM/>

5.2.2 RASTERAFBEELDINGEN

5.2.2.1 Officiële standaarden

5.2.2.1.1) *Tagged Image File Format (.tif, .tiff)*

Het TIFF-bestandsformaat 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 $Y_C B_C R_C$ -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, $Y_C B_C R_C$ en CIE Lab¹¹. TIFF wordt veel gebruikt bij de opslag van ingescande afbeeldingen en foto's. TIFF-bestanden worden 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 Packbitscompressie¹². 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 customisering. 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 5.5.2.2), TIFF voor PageMaker, Kodak TIFF, TIFF-documenten opgeslagen met Adobe Photoshop, enz. Programma's kunnen de TIFF-blokken negeren die ze niet begrijpen of kunnen zich beperken tot het inlezen van de blokken die ze nodig hebben. Dit biedt voor archivalistische 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 geleid tot een wildgroei van TIFF-bestanden met verschillende interne structuren die

¹¹ 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 grijs tinten 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 kleurstaalen. 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.

¹² Packbitscompressie is de run-length compressie die werd ontworpen door Apple.

enkel met één bepaalde toepassing kunnen worden ingelezen¹³. Hierdoor neemt de platformafhankelijkheid van het TIFF-formaat sterk af. Een echt platformafhankelijk 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.

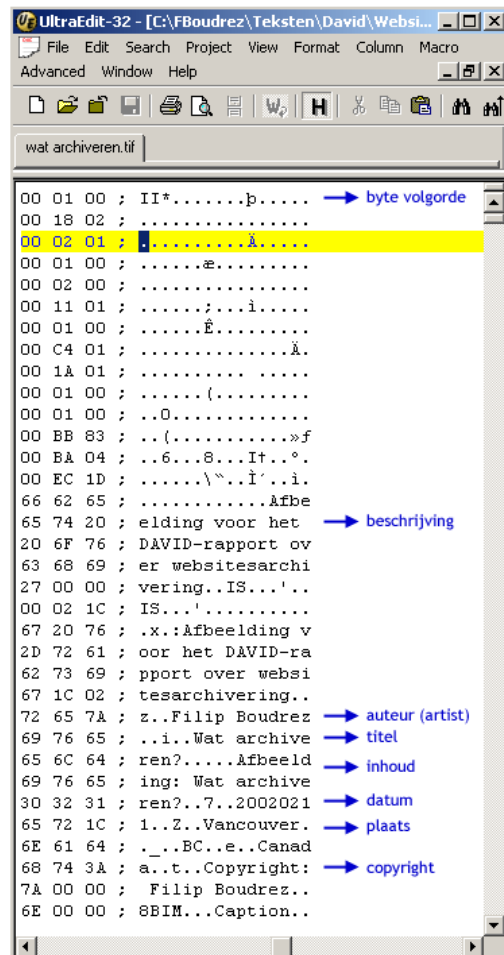
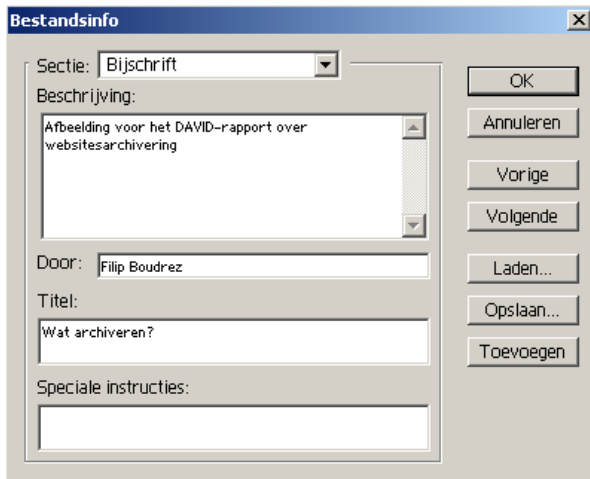
Een aantal velden van het TIFF-bestand zijn tekstblokken waarin metadata over de afbeelding wordt opgenomen. Deze velden bevatten geen binaire data, maar gewone ASCII- of Unicodekarakters. Zo kan men beschrijvende gegevens aan het TIFF-bestand toevoegen zodat ze er een essentieel onderdeel van worden. Het gebruiken van deze velden voor metadata brengt de platformafhankelijkheid 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
256	ImageWidth	short or long
257	ImageLength	short or long
258	BitsPerSample	short
259	Compression Schema	short (gebruikte compressie ¹⁴)
33432	Copyright	ASCII
306	DateTime	ASCII
316	HostComputer	ASCII (bijv. computer en OS die werd gebruikt bij de creatie van het bestand)
270	ImageDescription	ASCII-waarde
271	Make	ASCII (bijv. fabrikant van de scanner, video, apparatuur die werd gebruikt voor het maken van de afbeelding)
272	Model	ASCII (type scanner die werd gebruikt)
305	Software	ASCII (software naam en versienummer van de software die werd gebruikt voor het maken van de afbeelding)
TIFF: Extensions for document storage and retrieval		
269	DocumentName	ASCII
285	PageName	ASCII
297	PageNumber	ASCII

¹³ 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. MAC TAVISH, *DoD-NARA Scanned Images Standards Conference*, in: RLG Diginews, april 15, 1999, vol. 3, nr. 2.)

¹⁴ Het veldnummer 259 wordt gevolgd door een cijfer. Elk cijfer staat voor een bepaalde compressie
1: geen compressie5: LZW2: CCITT-groep 1 D6: JPEG3: CCITT-groep 332773: Packbits compressie4: CCITT-groep 4

Niet elke applicatie echter laat het invullen, editeren of raadplegen van deze metadata toe. Photoshop bijvoorbeeld laat de invulling van een aantal velden toe, maar de laatste versies promoten sterk het Adobe XMP metadatamechanisme. De metadata die in Photoshop via de documenteigenschappen worden ingevuld, worden bijgevolg soms als XMP metadata geregistreerd en niet als TIFF-tags. Een aantal producenten bieden (gratis of tegen betaling) programma's voor het inkapselen van TIFF-metadata aan, maar men kan evengoed gebruik maken van de gratis libtiff-bibliotheek. 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 4 en 5: 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.

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 bestandssysteemafhankelijk 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 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.

5.2.2.1.2) Joint Photographic Experts Group (.jpg, .jpeg)

Het JPEG-bestandsformaat is genoemd naar de groep experts (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.

JPEG - jFIF - SPIFF

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 gecompriemd 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, $(Y C_B C_R)$) en zwart-wit (grijschaal) foto's. JPEG ondersteunt geen transparantie.

De originele technische JPEG-specificatie schreef geen bestandsformaat voor data uitwisseling voor. Dit euvel werd opgelost door de ontwikkeling JFIF (JPEG File Interchange Format) in 1992. Dit is de officiële naam van het bestandsformaat, maar bijna iedereen gebruikt de benaming JPEG ipv JFIF.. Het derde deel van de standaard (uitgegeven in 1996) bevat wel een bestandsformaat: Still Picture Interchange File Format (SPIFF). SPIFF is complexer dan JFIF en ondersteunt buiten JPEG ook andere compressieschema's.

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.

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.

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 compressieprestatie 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>

5.2.2.1.3) Portable Network Graphics (.png)

PNG is ontworpen door het *World Wide Web Consortium* om een antwoord te bieden op de licentieproblemen 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 (LZ77). 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 compressiemethode 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 (interlacing), gamma correctie en twee dimensionale 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) zodat dezelfde afbeelding op verschillende computers er hetzelfde uitziet. PNG-bestanden bevatten eveneens een foutopsporings- en verbeteringsmechanisme zodat bijvoorbeeld transmissiefouten worden opgespoord. 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 werd in 2004 als een officiële standaard vastgelegd: ISO/IEC-15948 (*ISO/IEC 15948(2004): Information technology -- Computer graphics and image processing -- Portable Network Graphics (PNG): Functional specification*).

De verspreiding van het PNG-formaat kent een gestage opgang. PNG wordt vooral binnen internettoepassingen gebruikt. PNG is bijvoorbeeld het native bestandsformaat van Macromedia's Fireworks¹⁵. Naar schatting ondersteunen al bijna 200 verschillende computerprogramma's en een zestigtal webbrowsers het formaat. 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 apparaatafhankelijke kleurschema's.

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

5.2.2.2 Defacto standaarden

5.2.2.2.1) Bitmap (.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.

Vanwege hun platformafhankelijkheid worden BMP-afbeeldingen in de regel niet gebruikt als meest geschikt archiveringsformaat voor rasterafbeeldingen, al zijn hier wel uitzonderingen op mogelijk. Ze zijn evenmin bruikbaar als raadplegingsformaat, want hiervoor zijn de bestanden te groot.

Referentie: /

¹⁵ 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.

5.2.2.2.2) Graphics Interchange Format (.gif)

GIF is één van de oudste bestandsformaten voor afbeeldingen. De GIF-specificatie gaat terug tot 1987 en werd ontwikkeld door CompuServe Incorporated. 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 (1990) 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 wel 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 specificeren. 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 elke 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. De licenties voor het gebruik van grafische programma's die GIF-bestanden creëren, dekken de LZW-licentie 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 5.2.2.1.3) als alternatief gecreëerd. Niettegenstaande dit gegeven is GIF tot op de dag van vandaag een heel populair bestandsformaat in internet- en netwerktoepassingen. GIF-afbeeldingen maken bijvoorbeeld in grote getale deel uit van websitesarchieven.

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).

5.2.2.2.3) Encapsulated Postscript (.eps)

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 bitmapafbeelding 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

5.2.3 VECTORAFBEELDINGEN

5.2.3.1 Officiële standaarden

/

5.2.3.2 Defacto standaarden

5.2.3.2.1) Scalable Vector Graphics (.svg)

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. Inmiddels is specificatie 1.1 de huidige versie. Momenteel wordt al werk gemaakt van het vastleggen van SVG 1.2. 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 webbrowser¹⁶ (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 ingezoomd 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.

Naast archiveringsformaat voor afbeeldingen is SVG eveneens een potentieel archiveringsformaat voor Flash. SVG kan gebruikt worden voor de archivering van Flashobjecten die in webpagina's ingebed zijn en waarvoor een specifieke plug-in nodig is.

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

5.2.3.2.2) Drawing eXchange Format (.dxf)

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

¹⁶ <http://www.w3.org/Graphics/SVG/SVG-Implementations.htm>

recentste DXF-versie is 16.1.01. Door de grote wijzigingen is het DXF-formaat niet altijd compatibel met om het even welke AutoCAD versie.

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 aanduidt welk datatype wordt beschreven. De tags in een DXF-bestand zijn groepscode's. Bij elke nieuwe versie worden nieuwe groepscode's 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/>).

5.2.3.2.3) Drawing (.dwg)

DWG is het gesloten en eigendomsgebonden bestandsformaat van AutoCAD. Binnen de wereld van CAD/CAM-toepassingen heeft AutoCAD door zijn wijdverspreidheid de status verworven van standaard. DWG is een binair en heel compact bestandsformaat waarvan de specificatie door producent Autodesk niet wordt vrijgegeven. DWG ondersteunt 24 bits kleurdiepte, ware kleuren en 3-D modellen. Er bestaan dan ook nagenoeg (of helemaal geen?) andere applicaties dan AutoCAD die *.dwg-bestanden perfect 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. Als reactie op de monopoliepositie is de OpenDWG alliantie (nu Open Design Alliance) opgericht met als doel een open en gedocumenteerde industriële standaard voor CAD uitwisseling te creëren. Het OpenDWG-formaat is zo compatibel mogelijk met het DWG-formaat van Autodesk. De alliantie houdt zich ook bezig met het analyseren en ontleden van het DWG-formaat.

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 archiveringsformaat beschikbaar. Het OpenDWG formaat is wel publiek gedocumenteerd, maar het blijft echter afwachten of dit formaat ook voldoende ondersteuning en marktpenetratie heeft. 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 recentste AutoCAD-versie niet meer in staat is om bepaalde versies in te

lezen. Een andere mogelijkheid is het publiceren van de CAD-tekeningen als EPS-bestanden, ze in een rasterformaat of in PDF te bewaren.

Referentie: <http://www.opendwg.org/>

5.3 Audio

5.3.1 OFFICIËLE STANDAARDEN

5.3.1.1 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 MUSICAM-compressie; 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 en op joint stereo¹⁷.

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) standaard¹⁸. 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.

5.3.1.2 Pulse Code Modulation (.pcm)

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 WAVE en AIFF gebruiken de PCM-codec. De meeste audiocomputertoepassingen kunnen PCM inlezen. Bij de bits van een PCM-bestand staan de 1'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 (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., wat overeenkomst met ongeveer 1 gigabyte per uur.

¹⁷ Twee onafhankelijke kanalen: vb. 1 taal per kanaal

Joint stereo: efficiënte combinatie van twee kanelen (het linker en het rechter)

¹⁸ ISO/IEC 13818-7:1997 Information technology -- Generic coding of moving pictures and associated audio information -- Part 7: Advanced Audio Coding (AAC)

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 WAVE) geen fileheader 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 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 eenvoudige 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 Association. ADPCM bestaat ook in producentgebonden versies (Microsoft, Creative Labs, enz.).

PCM is de standaard audiocodec voor de digitale archivering van geluidsdocumenten. Vanwege de nood aan enkele minimale technische metadata over de audiodata is het aangewezen om PCM binnen een wrapperformaat (bijv. WAVE) te gebruiken.

5.3.2 DEFACTO STANDAARDEN

5.3.2.1 WAVE (.wav, .wave)

Het WAVE-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).

WAVE is Microsofts toepassing RIFF voor de opslag van audiobestanden. De WAVE-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 WAVE-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 WAVE-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. WAVE wordt bijgevolg soms als Windows PCM aangeduid. Samen met de wijdverspreidheid van Windowscomputers heeft dit voor gevolg dat ongecomprimeerde WAVE-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. WAVE-bestanden worden dan ook niet veel gebruikt binnen netwerktoepassingen. Voor uitwisseling over het internet worden WAVE-bestanden dan ook doorgaans naar MP3 omgezet of wordt binnen WAVE de MP3-codec gebruikt. Alle andere chunks in het WAVE-formaat zijn optioneel (cue, playlist, associated data, instrument, enz.).

Eén chunk biedt de mogelijkheid om metadata in het WAVE-bestand in te kapselen: de INFO-chunk. Deze chunk is optioneel. Applicaties die deze chunk niet ondersteunen, negeren de data in deze

chunk. De metadata worden als ASCII-waarden in het WAVE-bestand opgenomen. In deze INFO-chunk kunnen de volgende metadata worden geregistreerd:

- IARL - Archival Location: bijv. bestandsnaam of archiefnummer
- IART – Artist: bijv. naam uitvoerder(s)
- ICMT – Comments: commentaar, opmerking
- ICOP – Copyright: copyright informatie
- ICRD - Creation date: datum opname of digitalisering
- IENG – Engineer: naam ingenieur
- IGNR – Genre: genre
- IKEY – Keywords: trefwoorden (gescheiden door “;”)
- IMED – Medium: medium
- INAM - Name/Title: beschrijving/titel archiefdocument
- ISFT – Software: naam digitaliseringssoftware
- ISRC – Source: bron
- ISRF – Source Form: vorm/type origineel document
- ITCH – Technician: naam technicus

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

In een WAVE-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 WAVE-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.

WAVE kent een grote toepassing op zowel personal computers als in professionele opname-apparatuur. Het EBU (European Broadcast Union) heeft WAVE verder ontwikkeld tot BWF (Broadcast Wave Format) zodat uitwisseling tussen de verschillende Europese radiostations mogelijk is. EBU breidde de metadatavelden uit die in een WAVE-bestand kunnen worden ingekapseld en voorziet de volgende velden:

- Description: beschrijving (max. 256 karakters)
- Originator: naam producer (32 karakters)
- Originator Reference: informatie over de producer (32 karakters)
- Origination Date (yyyy-mm-dd): productiedatum
- Origination Time (hh:mm:ss): productietijd
- Time Reference : timecode van het audiobestand
- Coding History: historiek van het audiobestand, documenteren van de bewerkingen

Het archiveren van de digitale moederkopieën van geluidsdocumenten in WAVE-formaat biedt een aantal voordelen in vergelijking met archivering van geluid op audio-CD's

- WAVE-bestanden kunnen een grotere frequentie en bitdiepte hebben dan een audio-CD. De Red-Book standaard voor audio-CD's hanteert altijd 16 bits en 44,1 KHz. Een WAVE-bestand kan geluidsdocumenten met een bitdiepte tot 32 bits en een sample-rate tot 192 KHz bevatten.
- WAVE-bestanden kunnen meer metadata bevatten dan audiotracks op een audio-CD.
- archivering als WAVE-bestanden is mediumonafhankelijk. WAVE-bestanden kunnen op CD, tape of op harde schijf worden bewaard.

Het WAVE-formaat wordt binnen diverse digitaliseringsprojecten of archiveringscases als archiveringsformaat gebruikt (Phonogrammarchiv Wenen, IASA, cDAVID, enz.). Dit veronderstelt vanzelfsprekend dat binnen WAVE de uncompressed PCD-codec voor de audiodata wordt gebruikt.

Referentie: /

5.3.2.2 AU / SND (.au)

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 platformafhankelijk 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: /

5.3.2.3 Audio Interchange File Format (.aiff)

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 WAVE 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*

5.3.2.4 Free Lossless Audio Codec (.flac)

FLAC is een open source project met als doel een lossless audiocodec te ontwikkelen. De huidige versie is momenteel nummer 1.1.1 (oktober 2004). FLAC stelt niet alleen een codec, maar ook een

bestandsformaat, een API, voorbeeldsoftware, plugins voor players, enz. beschikbaar. De specificaties van de FLAC codec en het FLAC-formaat zijn open en vrij van patentrechten. FLAC-bestanden zijn platformonafhankelijk.

FLAC-bestanden zijn gecomprimeerd. In tegenstelling tot andere compressieformaten voor audio (MP3, OGG, enz.) is de FLAC-compressie lossless. Er treedt dus geen informatie- of kwaliteitsverlies op wanneer geluid dmv de FLAC-codec wordt opgeslagen. Na decompressie heb je opnieuw ongecomprimeerde PCM-geëncodeerde audiodata. FLAC ondersteunt lineaire PCM-geluidssignalen. Deze geluidssignalen kunnen variëren tussen 4 en 32 bits per sample en tussen 1 Hz en 655350 Hz.

Binnen een FLAC-bestand kan metadata worden bijgehouden. De FLAC-specificatie voorziet een aantal voorgedefinieerde metadatavelden, maar door het systeem van metadatatags kan men ook zelf metadatavelden toevoegen. De zelf gecreëerde metadatatags kunnen geregistreerd worden.

De integriteit van de FLAC-bestanden wordt gewaarborgd door de toepassing van CRC, MD5 en zogenaamde FLAC Fingerprints. Er worden twee soorten FLAC-bestanden onderscheiden: native FLAC en OGG FLAC. Bij native FLAC is de FLAC-audiodata verpakt in een FLAC-container, terwijl bij OGG FLAC de FLAC-audiodata in een OGG-container zit.

Voorbeelden waarbij FLAC wordt gebruikt voor archiveringsdoeleinden zijn niet bekend. Het valt ook te betwijfelen of FLAC bij de digitale archivering van geluid zal worden toegepast. Als raadplegingsformaat zijn FLAC-bestanden in vergelijking met MP3, RealMedia, OGG, enz. te groot (ca. 7 tot 9 MB/minuut digitaal geluid aan CD-kwaliteit). FLAC is door het gebruik van compressie evenmin een geschikt archiveringsformaat. Bovendien levert de lossless compressie maar relatief weinig winst op: FLAC-bestanden zijn ongeveer 10 tot 15 % kleiner dan ongecomprimeerde WAVE-bestanden. FLAC-compressie levert wel kleinere bestanden op dan Shorten (SHN).

Referentie: <http://flac.sourceforge.net/index.html>

5.3.2.5 OGG Vorbis (.ogg)

OGG Vorbis is een nieuw compressieformaat voor digitale audio. De naam OGG verwijst naar het Xiph.org's containerformaat voor audio, video en metadata. Vorbis is de naam van het compressieschema dat voor audio wordt gebruikt.

OGG Vorbis is volledig open en vrij van patentrechten. De Vorbis-codec werd ontwikkeld als antwoord op de vele bedrijfseigen, gepatenteerde codecs voor digitale audio. Volgens de Xiph.org Foundation behoort de Vorbis-specificatie volledig tot het publiek domein. De Vorbis-codec past lossy compressie toe. OGG Vorbis bestanden zijn bijgevolg slechts bruikbaar als raadplegingsformaat voor digitale audiodocumenten. OGG Vorbis is nog bijlange niet zo weidverspreid als MP3, maar begint wel aan populariteit te winnen. Ondertussen is er al in ruime mate hard- en software ondersteuning van het OGG Vorbisformaat.

OGG Vorbis is toepasbaar als raadplegingsformaat voor digitale audio. Het OGG-containerformaat is eventueel bruikbaar als archiveringsformaat, maar dan mag OGG niet in combinatie met de Vorbis-codec worden gebruikt. In de plaats van Vorbis wordt binnen OGG een lossless audiocodec gebruikt.

Referentie: <http://www.vorbis.com>

5.4 Video

5.4.1 OFFICIËLE STANDAARDEN

5.4.1.1 MPEG-Video (.mpeg, .mpg)

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 decompressiestandaarden voor de digitale opslag van video, audio en een combinatie van beide te creëren. De compressietechnieken zelf zijn veelal eigendom van de producenten.

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 1, 2 en 4 worden gebruikt voor computerbestanden die audio en/of visuele informatie bevatten. Deze 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 (zie 5.3.1.1) wordt de extensie *.mp gehanteerd. MPEG 7 en 21 zijn standaarden voor de beschrijving, de uitwisseling en het gebruik van multimedia inhoud.

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 tot 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"). MPEG-1 heeft een transmissiesnelheid van ongeveer 1,5 miljoen bits/seconde¹⁹.

MPEG-2²⁰ werd ontworpen voor digitale televisie en DVD aan uitzendkwaliteit (High Definition Television) met een transmissiesnelheid tussen 4 en 6 Mbps. MPEG-2 wordt hoofdzakelijk door de (digitale) televisie- en DVD-industrie gebruikt²¹. 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 op²². Als basis voor MPEG-4 werd het Quicktimeformaat gebruikt²³.

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.

¹⁹ 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*

²⁰ MPEG-3 was bedoeld voor High Definition Televisie (HDTV) toepassingen. MPEG-2 bleek echter aan de HDTV behoeften te voldoen zodat dit standaardiseringsinitiatief bij MPEG-2 werd ondergebracht.

²¹ MPEG-2: ISO/IEC 13818 (1994): *Information Technology--Generic Coding of Moving Pictures and Associated Audio*.

²² MPEG-4: ISO/IEC 14496 (1998): *Information technology -- Coding of audio-visual objects*

²³ <http://www.apple.com/pr/library/1998/feb/11iso.html>

MPEG-21 (2000) biedt een framework voor de uitwisseling en het gebruik van allerhande types multimedia²⁴.

Vanwege de toepassing van *lossy* compressie is MPEG niet geschikt als archiveringsformaat voor bewegend beeld, eventueel wel als raadplegingsformaat.

Referentie: <http://www.iso.ch>; <http://mpeg.telecomitalia.com/>; <http://www.mpeg.org>

5.4.1.2 Motion-JPEG2000 (.mj2, .mjp2)

Motion-JPEG2000 is geen echt bestandsformaat, maar is in de eerste plaats een codec voor de opslag van digitale video. Toch bestaat er ook een Motion-JPEG2000 bestandsformaat. Dit bestandsformaat is eigenlijk een ISO-Base Media bestandsformaat²⁵ waarbinnen de Motion-JPEG2000 codec wordt gebruikt.

Motion-JPEG2000 is gebaseerd op de JPEG2000-standaard. Met deze videocodec is het mogelijk om *lossless* waveletcompressie toe te passen op videobeelden. Ongecomprimeerde opslag van digitale video is vanuit archiveringsstandpunt meer aangewezen, maar vergt heel krachtige hard- en software voor de verwerking en opslag, wat (nog) niet altijd realiseerbaar is. *Lossless* compressie is in die gevallen de aangewezen optie. Voor archivering van digitale video wordt deze videocodec in combinatie met AAF en/of MXF gebruikt.

Motion-JPEG2000 compressie kan zowel *lossless* als *lossy* worden toegepast. Voor archiveringsdoelstellingen wordt *lossless* compressie toegepast. Dit levert een compressieratio van ongeveer 3:1 op. In tegenstelling tot bijv. MPEG2 en MPEG4 past Motion-JPEG2000 geen inter-frame compressie toe, maar enkel intra-frame compressie. Elk frame wordt als een afzonderlijk beeld opgeslagen en gecomprimeerd. Hierdoor zijn de frames individueel raadpleegbaar en editeerbaar. Momenteel zijn al commerciële implementaties beschikbaar.

Vanuit archiveringsperspectief zal Motion-JPEG2000 niet als bestandsformaat, maar als codec binnen een geschikt archiveringsformaat zoals AAF en MXF worden gebruikt.

Referentie: ISO/IEC 15444-3:2002/Amd.2:2003. Information technology -- JPEG 2000 image coding system -- Part 3: Motion JPEG 2000/Amendment 2: Motion JPEG 2000 derived from ISO base media file format

5.2.2 DEFACTO STANDAARDEN

5.2.2.1 Advanced Authoring Format (.aaf)

AAF is een bestandsformaat voor bewegend beeld en de bijhorende metadata. Een AAF bestand bevat de video-, de audio- en de metadata. AAF werd gecreëerd door de AAF Association. AAF is een non-profit groep die in 2000 officieel werd opgericht met als specifiek doel een uitwisselingsstandaard voor beeldmateriaal vast te leggen. De AAF-leden komen uit de softwareindustrie (Adobe, Microsoft, Apple,...), de televisieindustrie (Sony, Panasonic, Eastman Kodak, ...) en de televisiewereld (AOL/Time Warner (o.a. CNN en Warner Bros), BBC, FOX, ...). De AAF Association werkt samen met ISO en het Pro-MPEG Forum. Inmiddels is versie 1.1 van de AAF-specificatie vastgelegd.

²⁴ MPEG-21: ISO/IEC TR 21000-1(2001): Information technology -- Multimedia framework (MPEG-21) -- Part 1: Vision, Technologies and Strategy

²⁵ ISO/IEC 14496-12:2003. Information technology -- Coding of audio-visual objects -- Part 12: ISO Base Media File Format

Het initieel doel van de AAF Association was de ontwikkeling van een standaard voor de uitwisseling van de metadata in combinatie met (externe) video- en audiodata. Hierdoor kon voor editeren en postproductie verschillende softwarepakketten worden gebruikt zonder metadata te verliezen zodanig dat men voor elk type bewerking de beste software kan gebruiken. Uitwisseling van de video- en audio was met de gangbare systemen al mogelijk, maar dit ging telkens met metadata-verlies gepaard. Op die manier kunnen voortaan de grenzen van commerciële pakketten worden doorbroken. AAF kan echter evengoed als native file format voor essence (video- en audiodata) en metadata worden gebruikt. AAF is gebaseerd op Open Media Framework Interchange (OMFI), een standaard formaat voor digitale media uitwisseling ontwikkeld door Avid Technology.

De AAF standaard bevat:

- een object-georiënteerd model als basis voor de opslag van essence en metadata
- een open source en platform onafhankelijk SDK (software development kit) bestaande uit een API, Reference Implementation, een object manager en een opslagsysteem (Microsoft Structured Storage) dat op diverse platformen (MacOS, Unix varianten, Windows) en met verschillende netwerkprotocollen kan worden gebruikt. De API is geschreven in Interface Definition Language (IDL). Voor de programmeerinterface kan gebruik worden gemaakt van de COM-technologie (Component Object Model) van Microsoft. De SDK is beschikbaar op <http://aaf.sourceforge.net>
- voorbeeld software en demo bestanden

AAF kan meerdere types encoding voor video- en audiodata bevatten. AAF is met andere woorden niet gebonden aan één bepaald compressieschema (MPEG, DV). AAF kan net zo goed ongecomprimeerde videodata bevatten.

Het AAF-bestandsformaat is ook uitbreidbaar. Zowel voor de essence als de metadata zijn uitbreidingen in functie van specifieke noden mogelijk. Zo'n uitbreiding is bijvoorbeeld mogelijk om AAF als native file format te gebruiken. Met AAF wordt volledige tapeless opslag mogelijk. De uitwisseling en opslag kan volledig in netwerkgeving en op allerhande disks gebeuren.

AAF werd inmiddels al volop geïmplementeerd in professionele (editeer)software en is al commercieel beschikbaar. AAF wordt als professioneel uitwisselingsformaat voor (post)productie doeleinden gebruikt. De verwachting is dat AAF ook volop wordt ingepast in desktopsoftware voor het bewerken en bekijken van bewegend beeld. Momenteel lopen ook de eerste onderzoeksprojecten waarin de bruikbaarheid van AAF voor archiveringsdoeleinden wordt onderzocht. Op het eerste zicht is AAF hiervoor inzetbaar. Alle programma elementen kunnen op een consistente wijze voor archivering worden verpakt. Wel dient men erop te letten dat alle data in het AAF-bestand wordt ingekapseld en dat er niet naar extern video- of audiodata wordt gerefereerd. Een interessante eigenschap vanuit archiveringsperspectief is dat de volledige historiek van bron tot eindproductie van elke programma element kan geregistreerd worden.

AAF is ondertussen al geïmplementeerd in diverse commerciële producten: Windows Media Video en Audio 9 Series, Windows Media Player (Microsoft, plugin ontwikkeld door *Colledia*-project), Xpress Pro - Mojo (Avid), Première Pro (Adobe), Final Cut Pro (Apple), enz.

AAF is niet in de eerste plaats een afspeel- en/of streamingformaat. Hiervoor werd MXF ontwikkeld (zie onder). AAF en MXF zijn complementair. Op basis van AAF-bestanden kunnen gemakkelijk MXF-bestanden, maar even goed ook andere outputformaten, worden samengesteld.

Referenties: <http://aafassociation.org>; B. GILMAR, *AAF-The Advanced Authoring Format*, in: *EBU Technical Review*, juli 2002; B. TURNER, *AAF - 'The Super-EDL'*, in: *Broadcast Engineering Magazine*, mei 2002.

5.2.2.2 Material Exchange Format (.mxf)

Het MXF-bestandsformaat voor bewegend beeld is het resultaat van de samenwerking tussen het Professional MPEG Forum en de AAF Association (zie boven). Het Pro-MPEG Forum is een

vereniging van broadcasters, programmamakers en producenten ter promotie van de MPEG-2 standaard. MXF werd voor standaardisatie voorgelegd aan SMPTE. MXF wordt ondersteund vanuit de industrie (MOG Solutions, Sony die een MXF SDK ontwikkelde, enz.), de open source gemeenschap (Open Source MXF SDK), enkele belangrijke omroepen (o.a. BBC), enkele onderzoeksinstituten (Institut für Rundfunktechnik) en bepaalde gebruikersgroepen zoals EBU.

MXF is een subset van AAF. MXF hergebruikt een afgeslankte versie van het AAF object model. Voor video en audio data worden van het AAF object model enkel de delen mbt de rushes en de afgewerkte programma's overgenomen. Objecten, attributen en methoden mbt composities, effecten en de in-file dictionary ontbreken in het MXF object model. Dit heeft voor gevolg dat de binaire formaten van MXF en AAF grondig verschillen, dat MXF een minder complexe bestandsstructuur heeft en dat MXF-bestanden een kleinere bestandsomvang hebben waardoor ze meer geschikt zijn voor on line access. Ook inzake metadata hanteert MXF een beperkter set dan AAF.

MXF is specifiek ontwikkeld voor de raadpleging van afgewerkt videomateriaal dat (in streaming) vanop videoservers beschikbaar worden gesteld. De inhoud van MXF-bestanden kan gemakkelijk in streaming worden omgezet. De inhoud van MXF-bestanden wordt in principe niet meer bewerkt. In tegenstelling tot AAF zijn MXF-bestanden altijd volledig en zelfvoorzienig en bevatten ze geen verwijzingen naar externe bronnen.

Eén MXF-bestand bevat zowel de video- en audiodata als de metadata. Voor de metadataset, -handling en -opslag wordt gebruik gemaakt van XML.

De body bevat een sequentie van videoframes, met audio en frame-based metadata. Deze metadata bevat timecode en informatie over het bestandsformaat (interleaved mediafile). Net zoals AAF is de MXF body niet compressie of bit rate afhankelijk. MXF-bestanden kunnen dus net zo goed niet, *lossless* of *lossy* gecomprimeerd zijn. Dit hangt af van de videocodec die wordt gebruikt. Een veelgebruikte *lossless* videocodec is Motion-JPEG2000. Migratie van AAF naar MXF is *lossless* wanneer geen ander compressieschema wordt toegepast.

MXF en AAF zijn complementair. Niet alleen kan men op basis van AAF-bestanden gemakkelijk MXF-bestanden samenstellen, MXF video en audio wordt soms ook gecombineerd met AAF metadata. Een mogelijke toepassing van AAF en MXF voor digitaal videomateriaal is de moederkopie bewaren als AAF-bestanden en voor de raadpleging MXF-bestanden verspreiden.

Referenties: <http://www.pro-mpeg.org>; <http://aafassociation.org>

5.2.2.3 Audio Video Interleave (.avi)

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 bestandsomvang te verkleinen. Hierbij gaan frames verloren en kan de slow motion functionaliteit wegvallen.

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.

Vanwege de platformafhankelijkheid en het gebruik van compressie is AVI geen geschikt archiveringsformaat voor video. AVI is daarentegen wel bruikbaar als raadplegingsformaat (bijv. verspreiding via internet).

5.2.2.4 Quicktime (.qt, .mov)

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 recentste 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 gecompriemd. 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, WAVE, 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.

Net zoals AVI is Quicktime niet geschikt als archiveringsformaat, maar eventueel wel als raadplegingsformaat.

Referentie: <http://developer.apple.com/documentation/QuickTime/QTFF/qtff.pdf>

5.2.2.5 Flash (.swf)

SWF of voluit Shockwave Flash wordt ontwikkeld en beheerd door producent Macromedia. Het SWF-bestandsformaat (uitgesproken als "swif") wordt gebruikt om animaties met (raster en/of vectoriële) afbeeldingen, video en geluid op het WWW te verspreiden. SWF wordt ook gebruikt voor dynamische navigatiemenu's, banners en buttons in websites. De animaties worden geprogrammeerd met Macromedia's ActionScript en zijn gekoppeld aan een tijdlijn. Macromedia publiceert de SWF-specificatie zodat andere producenten ook tools voor het samenstellen en bekijken van SWF-bestanden kunnen ontwikkelen. In dit opzicht is de standaardiseringsstatus van SWF vergelijkbaar met die van PDF van Adobe Corporation (open, producentgebonden, defacto standaard). Momenteel is SWF 7.0 de huidige versie van het bestandsformaat.

Enigszinds verwarrend wordt SWF het Flash bestandsformaat genoemd. SWF wordt hierdoor dikwijls verward met het bestandsformaat waarin de applicatie Flash van Macromedia zijn documenten of projecten bewaard. Deze documenten hebben echter de extensie ".fla". SWF is het bestandsformaat waarin animaties ontwikkeld in Flash-projecten worden gepubliceerd. SWF is met andere woorden het raadplegingsformaat, en niet het editeerformaat (= ".fla"). Dit neemt echter niet weg dat SWF-

bestanden niet meer bewerkt kunnen worden. SWF-bestanden zijn wel editeerbaar maar moeten voorafgaand gedecompileerd worden.

Voor het raadplegen van SWF-animaties is een Flash-player nodig. Macromedia verspreidt zelf een player, maar men kan evengoed de player van een andere producent gebruiken.

SWF-animaties zijn meestal zelf-voorzienig, al kunnen ze ook afhankelijk zijn van externe bronnen (bijv. video). SWF past volop compressie toe zodat de SWF-bestanden relatief snel uitwisselbaar zijn. Toch zijn SWF-bestanden veel groter dan bijv. HTML-pagina's met bijhorende afbeeldingen en stylesheets.

Strikt genomen wordt SWF niet als een "natuurlijke webtechnologie" zoals (X)HTML of XML beschouwd. SWF is een binair bestandsformaat en is bijgevolg niet "menselijk leesbaar". Dit neemt niet weg dat heel veel websites in SWF zijn of veel SWF-animaties bevatten. SWF is het zevende meest voorkomende bestandsformaat in websitesarchieven. Tegenstanders van Flash stellen Scalable Vector Graphics (SVG) – gebaseerd op XML en dus wel een "native" webtechnologie – als alternatief voor websites met animaties voor. SVG wordt soms ook als archiveringsformaat voor SWF-bestanden voorgesteld, maar het lijkt twijfelachtig of SVG alle essentiële functionaliteiten van SWF kan overnemen.

Vanuit perspectief van lange termijn raadpleegbaarheid is het binaire karakter van SWF-bestanden geen onoverkomenlijk probleem. De SWF-specificatie is immers gepubliceerd zodat het in principe ten alle tijde mogelijk is om een viewer of een player te creëren. De inkapseling van hyperlinks in SWF-bestanden daarentegen kan bij de archivering van websites wel problemen vormen:

- Het maken van een momentopname of een snapshot van een website is één van de methoden om een website te archiveren. De bijzondere tool die hiervoor wordt gebruikt (een webharvester of webcrawler) moet in staat zijn om de ingebedde hyperlinks te extraheren zodat ook de gelinkte webpagina's kunnen vastgelegd worden. Er zijn echter nog maar weinig webharvesters of webcrawlers die met deze functionaliteit uitgerust zijn.
- De ingekapselde absolute links dienen omgezet te worden naar relatieve links, zodat de gebruiker bij raadpleging binnen de gearchiveerde website blijft. De hyperlinks binnen een SWF-bestand zijn niet zomaar editeerbaar. Idealiter beschikt men hiervoor over het ".fla"-bronbestand, op basis waarvan men vervolgens een nieuw SWF-bestand publiceert. Bij gebrek aan het bronbestand kan men ook het SWF-bestand decompileren, maar dit is niet de meest aangewezen weg (bescherming van het auteursrecht, SWF-bestanden zijn doorgaans beveiligd met paswoorden, enz.)

Referentie: http://download.macromedia.com/pub/flash/flash_file_format_specification.pdf; R. Entlich, *Flash in the Pan or Around for the Long Haul? Assessing Macromedia's Flash Technology*, in: *RLG-DigiNews*, juni 2004 (http://www.rlg.org/en/page.php?Page_ID=17661#article3).

5.5 Geografische informatie

5.5.1 OFFICIËLE STANDAARDEN

/

5.5.2 DEFACTO STANDAARDEN

5.5.2.1 Geography Markup Language (GML)

Geogerefererde gegevens worden al jarenlang in gesloten en producentgebonden formaten opgeslagen, met uitwisselingsproblemen en software- en versiegebondenheid als gevolg. Voor het beheer, de analyse en de visualisering van deze informatie was bijgevolg specifieke GIS- of CAD-software vereist. Officiële internationale standaardiseringsinitiatieven kenden geen positief gevolg. Dit

leidde tot geïsoleerde standaardisatie-initiatieven in verschillende landen (bijv. SDTS in USA²⁶). Commerciële spelers probeerden deze marktlanune in te vullen. Enkele commerciële formaten groeiden uit tot defacto standaarden (bijv. shapebestand en E00-formaat van ESRI) en waren tot op beperkte hoogte bruikbaar voor de uitwisseling van geospatiale informatie.

In 1994 werd het OpenGIS Consortium (OGC) in het leven geroepen om de incompatibiliteitsproblemen binnen de GIS-gemeenschap op te lossen. Het OGC had als doel een oplossing te zoeken voor de uitwisseling, de integratie en de opslag van geogerefererde data. Het OGC richtte zich in de eerste plaats tot het zoeken naar gemeenschappelijke interfaces (OpenGIS Interfaces) en protocollen. Met het vastleggen van de XML-specificatie beschikte het OGC over een basis encoderingswijze voor GIS-data. Uit het huwelijk van W3C's XML en OpenGIS Interfaces ontstond binnen de schoot van OGC vervolgens een nieuw gemeenschappelijk geodataformaat: Geography Markup Language (GML). GML werd getest tijdens de OpenGIS Consortium Web Mapping Test Bed (september 1999).

GML is gebaseerd op het gemeenschappelijk geografisch model van OGC: de OpenGIS Abstract Specification. GML is een XML encoding voor de weergave van geografische objecten volgens het OGC geografisch model en heeft twee belangrijke standaarden als uitgangspunt: het Open GIS geografisch model en XML. Volgens dit abstract model bestaat geografie uit geografische objecten ('features'²⁷ genoemd). De specificatie bevat standaarden voor het weergeven van features en verzamelingen features. In een GML-bestand worden (verzamelingen) features beschreven. De beschrijving van een feature is niets meer dan een opsomming van zijn eigenschappen en zijn geometrie. De eigenschappen hebben telkens een naam, een type en een waarde. Welke eigenschappen een bepaalde feature heeft, wordt door zijn definitie voorgeschreven. De geometrie is de vorm en lokalisering van het ruimtelijk object. Punten, lijnen, curven, oppervlakken en polygonen zijn de geometrische basisvormen. Elk feature kan meerdere features bevatten. Het feature 'park' bestaat bijvoorbeeld uit de features 'boom', 'pad', 'vijver', 'bank', 'vuilnisbak', 'lantaarnpaal', enz.

Op 12 mei 2000 werd GML versie 1.0 vastgelegd. Deze versie bestond uit drie profielen (GML.1.0, GML.2.0 en GML.3.0.²⁸) en was gebaseerd op een combinatie van meerdere DTD's en RDF. Vanwege de algemene tekortkomingen van DTD's (beperkte validatiemogelijkheden, geen ondersteuning van namespaces, geen type inheritance, enz) was er binnen GML 1.0 nood aan RDF om deze hiaten in te vullen. Tot op heden vond RDF echter geen algemene ingang. Op 20 februari 2001 werd vervolgens GML 2.0 goedgekeurd. GML 2.0 werd volledig op XML Schema gebaseerd. XML Schema had ten tijde van het samenstellen van GML 2.0 inmiddels de status van W3C Recommendation gekregen. XML Schema vult eveneens de DTD hiaten in en kent in tegenstelling tot RDF wel een grote verspreiding. De GML-specificatie schrijft eveneens een welbepaald ruimtelijk referentiesysteem voor. GML 2.0 had binnen het OGC de status van 'Implementation Specification'. De GML-specificatie 2.0 bevat drie basisschema's: feature.xsd, geometry.xsd en xlink.xsd. De eerste twee schema's bepalen hoe features en hun geometrie in GML worden vastgelegd. Het derde schema xlink legt vast hoe geografische data kunnen gekoppeld worden aan externe data. GML 2.0 is aanvaard door UK Ordnance Survey (bijv. MasterMap) en GeoMatics Canada. GML wordt momenteel door een aantal producenten (bijv. Esri, Galdos System Inc.) in hun software geïmplementeerd. Voorbeelden van GML 2.0 documenten zijn beschikbaar op de website van de Technische Universiteit Delft²⁹. Belangrijke beperkingen van GML 2.0 waren: enkel ondersteuning van 2D geometrie en lineaire elementen, geen topologie, geen versiebeheer, geen temporele attributen. Ontwikkelaars konden deze beperkingen wel omzeilen door hiervoor in hun eigen applicatieschema's oplossingen te voorzien, maar het bleef moeilijk om reële geografische applicaties integraal in GML weer te geven.

Inmiddels is versie 3 van GML voltooid. De specificatie werd op 29 januari 2003 vastgelegd en is vrij beschikbaar op de website van het OpenGIS consortium. Terwijl versie 2.0 zich nog hoofdzakelijk op simple features (twee dimensionele coördinaten en lineaire interpolatie) richtte, biedt versie 3.0 de mogelijkheid om complexe geografische features in GML te beschrijven. Versie 3.0 ondersteunt:

²⁶ <http://mcmweb.er.usgs.gov/sdts/>; <http://data.geocomm.com/sdts/>

²⁷ Het OGC omschrijft een 'feature' als 'A feature is an abstraction of a real world phenomenon; it is a geographic feature if it is associated with a location relative to the earth'.

²⁸ GML.1.0: vaste DTD's; GML.2.0: vaste DTD's + applicatie specifieke DTD's, GML.3.0: RDF en RDF Schema.

²⁹ http://www.gdmc.nl/GML-relay/gml_relay_13_december_2002_in_em.htm

- complexe features: niet-lineaire interpolatie, 3D geometrie, 2D topologie, dynamische features, features met tijdelijke of complexe eigenschappen, enz.
- topologie of relaties tussen features
- referentiesystemen voor ruimte en tijd, meeteenheden en info over standaarden

Versie 3 biedt eveneens een aantal voorgedefinieerde stijlen voor de visualisatie van features en coverages en is conform andere geografische standaarden (o.a. ISO DIS 19107, 19108, 19118, 19123). Om dit allemaal mogelijk te maken werden de basisschema's van GML versie 2.0 uitgebreid. De basisschema's van versie 3.0 zijn 8 keer zo lang en bevatten meer dan verschillende 1000 GML-elementen. Vanwege de gebruiksvriendelijkheid is het mogelijk om GML versie 3.0 modulair te gebruiken, zodat de gebruikers of ontwikkelaars de componenten kunnen selecteren die ze effectief wensen te gebruiken binnen hun geografische toepassing. Op die manier worden als het ware subsets van GML 3.0 gemaakt. Het OpenGIS verwacht dat er per IT-domein gemeenschappelijke GML-profielen worden gemaakt. Versie 3.0 is achterwaarts compatibel met versie 2.0, maar niet voor de volle 100 %. Een aantal schema componenten van GML 2.0 zijn 'deprecated'.

De ontwerpschema's van GML 3.0 werden ingediend bij ISO/TC 211³⁰ met de bedoeling GML 3.0 als internationale standaard vast te leggen (projectnr. 19136) zodat GML kan gebruikt worden als opslagformaat binnen GIS of als uitwisselingsformaat tussen verschillende GIS-systemen. De ontwerpschema's van GML 3.0 werden afgestemd op G-XML³¹, een Japans gelijkaardig standaardiseringsinitiatief voor de uitwisseling van geografische informatie. Versie 3 heeft eveneens de standaard ISO-19107 als geometrisch en topologisch model aangenomen. Standaard ISO-19115 wordt voor de geografische metadata gebruikt. Inmiddels werd versie 3.0 al verfijnd tot versie 3.1. Deze laatste versie bevat enkele nieuwe geometrieën en is verder afgestemd op andere internationale standaarden.

GML gebruikt de XML syntax als noteringwijze voor digitale ruimtelijk gerefereerde informatie. Doordat GML op XML is gebaseerd, erft GML een aantal voor- en nadelen van XML: uitbreidbaar, (semi-) gestructureerde data opslag, scheiding van presentatie en inhoud, geschikt voor uitwisseling en opslag, integratie met andere (types) bronnen mogelijk (bijv. bevolkingsgegevens, nationale statistieken) en natuurlijk digitale duurzaamheid. GML beantwoordt hierdoor aan de vraag voor integratie tussen GIS-systemen onderling en eventueel met andere informatiesystemen binnen de organisatie. Zeker wanneer de informatie in XML aanwezig is, ligt een integratie met geogerefereerde data voor de hand. Externe data of attributen kunnen gerelateerd worden aan de geografische features dmv xlink of kunnen in de GML-documenten zelf worden ingekapseld als attributen van de GML-objecten.

GML baseren op XML houdt in dat de XML-syntax wordt toegepast bij het registreren van geografische informatie in GML-documenten. GML is een specificatie waarin onder meer een basisset van ruimtelijke objecten, hun onderlinge relaties en een gemeenschappelijk datamodel zijn vastgelegd. In die zin kan GML met een XML namespace worden vergeleken, waarbij een set van vaste tags wordt gebruikt voor de beschrijving van geospatiale data en een variëteit aan geometrische types.

Aangezien GML de uitbreidbaarheid van XML overerft, kan een gebruiker(sgroep) zijn eigen datamodellen in zijn eigen GML-formaat creëren. Hiertoe wordt een eigen XML Schema vastgelegd, waarbij de basisfeatures van de GML-specificatie verder worden verfijnd of aangepast aan de eigen specifieke noden. Naast de basiselementen biedt de GML-specificatie immers een geheel van regels en richtlijnen aan voor het uitwerken van een eigen applicatiedocumentmodel met behulp van het W3C XML Schema. Zo kan men in zijn eigen applicatieprofiel de eigenschappen van bepaalde basisfeatures verder uitbreiden. Het XML Schema definieert het lokaal GML-profiel. GML-documenten bevatten immers nieuwe types features, collecties, eigenschappen of geometrische eigenschappen die in de meeste gevallen aanpassingen of uitbreidingen zijn van de basistypen die in de GML specificatie zijn vastgelegd. Dit past perfect in de object-georiënteerde benadering die met GML wordt nagestreefd. Op basis van de prefix in de XML-tag kan men achterhalen of er wordt verwezen naar het

³⁰ <http://www.isotc211.org>

³¹ G-XML is een project van het Database Promotion Center (Japan) en heeft tot doel een protocol uit te werken voor de uitwisseling van geografische informatie via het internet op basis van XML. De laatste versie is 2.5 (<http://gisclh01.dpc.or.jp/gxml/contents-e/>)

basistype van de GML standaard of naar een lokaal profiel. Het gebruik van ad hoc applicatieschema's heeft als nadeel dat het uitwisseling bemoeilijkt. De uitwisseling van de applicatieschema's is onontbeerlijk zodat softwaremodules externe schema's kunnen interpreteren. Gewone XML parsers kunnen de structuur van applicatieschema's ontleden zodat applicaties op een dynamische wijze kunnen achterhalen welk XML element binnen het GML document een feature, een eigenschap en geometrische eigenschappen weergeeft. GML software kan in principe met meerdere XML Schema's werken.

De wijze waarop de ruimtelijke objecten worden gevisualiseerd is geen onderdeel van GML-documenten. GML versie 3 voorziet wel een voorgedefinieerde voorstellingswijze van de GML-objecten, maar binnen het GML-concept worden de geografische data en de visuele voorstellingen immers van elkaar gescheiden. De GML-bestanden bevatten enkel de data over de eigenschappen en de geometrie van de geogerefererde entiteiten. Hoe deze entiteiten visueel worden voorgesteld, is niet ingekapseld in de GML-bestanden zelf. De GML-data kunnen wel op een grafische wijze worden voorgesteld. In de meeste gevallen is een kaart de grafische interpretatie van de data. De grafische formaten gebaseerd op XML zijn hiervoor geschikt. Met SVG (tweedimensioneel) en Extensible 3D (X3D: driedimensioneel) beschikt men over op XML gebaseerde grafische formaten. De omzetting van GML naar SVG of X3D (eventueel nog VML, VRML) gebeurt op basis van een XSLT-transformatie. XSLT is de geschikte technologie voor de transformatie van een data encodingstaal naar een presentatietaal. SVG en X3D sluiten goed aan bij GML.

Op hetzelfde GML-bestand kunnen meerdere stylesheets worden toegepast. De interpretatie en visualisering van de GML kan binnen een server-clienttoepassing volledig aan de clientzijde worden uitgevoerd zodat de gebruiker zijn eigen voorkeuren en stylesheets kan toepassen ('custom map styling': bijv. aanduiden van hoofdwegen in rode stippellijn). Voor de grafische voorstelling van geodata in kaarten of plattegronden heeft men bijgevolg geen bijzondere software nodig. Een recente webbrowser, eventueel uitgebreid met een SVG plug-in (bijv. Adobes SVG viewer), volstaat. Hierdoor worden problemen bij het uitwisselen van vectoriële afbeeldingen vermeden. In de plaats van een vectoriële afbeelding uit te wisselen waarvoor specifieke software was vereist, worden de GML-bestanden op basis waarvan een vector wordt samengesteld aan de client bezorgd. Uit dit alles volgt dat GML zich op de eerste plaats richt op de data van GIS-toepassingen (eigenschappen en geometrie van grafische objecten) en in mindere mate op de kaarten die binnen GIS-omgevingen worden gecreëerd.

De visualisering van GML-documenten in de vorm van kaarten is overigens maar één van de outputmogelijkheden. Geografische objecten vastgelegd in GML-documenten kunnen even goed als een tekstdocument of als gesproken woorden worden weergegeven. Door de scheiding van presentatie en inhoud beantwoorden de GML-bestanden aan de noden van multi-purpose of multi-channel publishing op basis van één gemeenschappelijke bron. De ontwikkeling van het WWW en het mobiele internet doet GML nog aan belang winnen. GML maakt data uitwisseling via het web voor verschillende doeleinden en outputvarianties mogelijk. Net zoals XML kunnen GML-documenten gevalideerd worden met gewone parsers en geëditeerd worden met diverse computerprogramma's (teksteditors, CAD en GIS software).

GML kan op diverse wijzen in de praktijk worden toegepast. In zijn meest doorgedreven toepassing worden de geografische data rechtstreeks als GML-bestanden binnen GIS bijgehouden. Het nadeel hiervan is natuurlijk de grote omvang. Geografische datasets hebben op zich al een grote omvang en bewaring als GML-bestanden met hun tekstuele encoding en tags doet dit volume nog toenemen. Versie 3.0 zou hiervoor een oplossing bieden. Binnen bestaande GIS-toepassingen wordt GML vooral als uitwisselingsformaat gebruikt. Met GML wordt geografische data op het internet verspreid of tussen niet compatibele systemen uitgewisseld. GML-documenten kunnen op meerdere wijzen worden gegenereerd vanuit relationele of object-georiënteerde databanken: gebruik van JAVA componenten, bevraging van de databanken dmv XSL(T) en samenstellen van de GML-documenten door een web feature server (WFS), enz. Met dezelfde tools kunnen eveneens GML-documenten aan databanken worden toegevoegd³². GML data kunnen ook rechtstreeks als XML-documenten in een nativ XML database worden bijgehouden (bijv. Cartagena, Tamino, Oracle XML database).

³² Bijv. GO Loader van Snowflake software kan data uit een GML-document in een Oracle 9i database laden.

Implementatievoorbeelden van webservices gebaseerd op GML zijn onder andere: IONIC (GML via WFS), TUDelft (GML via WFS), Esri NL (Shape via WGS en Web Image Service). In de meeste GML-toepassingen wordt een XML laag rond bestaande applicaties gebouwd.

GML is dus inzetbaar als uitwisselingsformaat tussen niet compatibele databanksystemen of als duurzaam archiveringsformaat. Binnen een GIS-proces kan GML gebruikt worden vanaf de dataverwerking, via opslag en verwerking (analyse) tot de verspreiding. Alleen voor de visualisatie is bijvoorbeeld SVG of VML nodig.

Volgens ESRI - toch één van de belangrijkste GIS-producenten - zijn een aantal nadelen aan GML verbonden. GML is niet performant genoeg vanwege de Unicode-encoding van de data en kan niet alle GIS-functionaliteiten ondersteunen. GML-voorstanders argumenteren dat GML compressie het performantieprobleem kan opvangen en dat het slechts een kwestie van tijd is alvorens GML alle functionaliteiten biedt.

Voorbeelden van archivering van GIS-data als GML-bestanden zijn nog niet bekend. Voor archivering biedt GML volgende voordelen:

- combinatie van data modellering, encoding, opslag en transport
- archivering in een gestandaardiseerd bestandsformaat (waarschijnlijk binnenkort zelfs officieel gestandaardiseerd door ISO).
- scheiding van data en visualisatie: GML is data georiënteerd
- GML-bestanden zijn in hoge mate platformafhankelijk en zelfbeschrijvend: GML kunnen door verschillende computersystemen worden geïnterpreteerd en zijn in hoge mate begrijpbaar voor mensen
- archivering van geospaatial data als tekstbestanden
- XML-functionaliteiten: bewaring van data samen met zijn betekenis, inkapseling of koppeling van attributen en metadata, validatiemogelijkheden, platformafhankelijk, integratie met andere informatiesystemen of niet-ruimtelijke data, gemakkelijk transformeerbaar, enz.
- vlot integreerbaar met gerelateerde archiefdocumenten in XML
- bruikbaar in combinatie met andere XML-technologieën
- aanpasbaar aan concrete archiefnoden

Referenties: <http://www.opengis.org>; <http://www.opengis.org/techno/documents/02-023r4.pdf>;

5.5.2.2 GeoTIFF (.tif, .tiff)

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 platformafhankelijk, 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. Hiervoor 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>

6. UITGEBREIDE INHOUDSOPGAVE

0. INHOUDSOPGAVE.....	1
1. BELANG VAN STANDAARDEN VOOR DIGITALE ARCHIVERING.....	1
2. HIËRARCHIE VAN DE ARCHIVERINGSSTANDAARDEN.....	2
3. GESCHIKTE ARCHIVERINGSFORMATEN.....	3
4. CODETABELLEN.....	3
4.1. OFFICIËLE STANDAARDEN.....	4
4.1.1 ASCII of ISO-646.....	4
4.1.2 ISO/IEC-8859.....	4
4.1.3 ISO-10646 en UNICODE.....	5
4.1.4 Andere ISO-codetabellen.....	5
4.2. DEFACTO STANDAARDEN.....	6
4.2.1 Unicode.....	6
4.2.2 EBCDIC.....	6
4.2.3 Base64.....	6
5. BESTANDSFORMATEN.....	7
5.1 TEKSTDOCUMENTEN.....	8
5.1.1 <i>Officiële standaarden</i>	8
5.1.1.1 Platte tekstbestanden (.txt).....	8
5.1.1.2 Standard Generalized Markup Language (.sgml).....	9
5.1.1.3 HyperText Markup Language (.htm, .html): 4.01.....	10
5.1.1.4 Open Document Architecture (.oda) and Interchange Format.....	11
5.1.2 <i>Defacto standaarden</i>	11
5.1.2.1 eXtensible Markup Language (.xml).....	11
5.1.2.2 Tagged Image File Format (.tif, .tiff).....	13
5.1.2.3 HyperText Markup Language (.htm, .html).....	13
5.1.2.4 PostScript (.ps).....	14
5.1.2.5 Portable Document Format (.pdf).....	14
5.1.2.6 OpenOffice XML - OpenDocument.....	19
5.1.2.7 Rich Text Format (.rtf).....	20
5.1.2.8 MS Word (.doc).....	20
5.1.2.9 ARC (.arc).....	21
5.2 AFBEELDINGEN.....	22
5.2.1 <i>Afbeeldingen in meta-formaat</i>	23
5.2.1.1 Officiële standaard.....	23
5.2.1.1.1 Computer Graphics Metafile (.cgm).....	23
5.2.2 <i>Rasterafbeeldingen</i>	24
5.2.2.1 Officiële standaarden.....	24
5.2.2.1.1 Tagged Image File Format (.tif, .tiff).....	24
5.2.2.1.2 Joint Photographic Experts Group (.jpg, .jpeg).....	28
JPEG - jFIF - SPIFF.....	28

JPEG-LS.....	29
JPEG-2000.....	29
5.2.2.1.3) Portable Network Graphics (.png).....	30
5.2.2.2 Defacto standaarden.....	31
5.2.2.2.1) Bitmap (.bmp).....	31
5.2.2.2.2) Graphics Interchange Format (.gif).....	32
5.2.2.2.3) Encapsulated Postscript (.eps).....	33
5.2.3 Vectorafbeeldingen.....	34
5.2.3.1 Officiële standaarden.....	34
5.2.3.2 Defacto standaarden.....	34
5.2.3.2.1) Scalable Vector Graphics (.svg).....	34
5.2.3.2.2) Drawing eXchange Format (.dxf).....	34
5.2.3.2.3) Drawing (.dwg).....	35
5.3 AUDIO.....	36
5.3.1 Officiële standaarden.....	36
5.3.1.1 MPEG-Audio.....	36
5.3.1.2 Pulse Code Modulation (.pcm).....	36
5.3.2 Defacto standaarden.....	37
5.3.2.1 WAVE (.wav, .wave).....	37
5.3.2.2 AU / SND (.au).....	39
5.3.2.3 Audio Interchange File Format (.aiff).....	39
5.3.2.4 Free Lossless Audio Codec (.flac).....	39
5.3.2.5 OGG Vorbis (.ogg).....	40
5.4 VIDEO.....	41
5.4.1 Officiële standaarden.....	41
5.4.1.1 MPEG-Video (.mpeg, .mpg).....	41
5.4.1.2 Motion-JPEG2000 (.mj2, .mjp2).....	42
5.4.2 Defacto standaarden.....	42
5.4.2.1 Advanced Authoring Format (.aaf).....	42
5.4.2.2 Material Exchange Format (.mxf).....	43
5.4.2.3 Audio Video Interleave (.avi).....	44
5.4.2.4 Quicktime (.qt, .mov).....	45
5.4.2.5 Flash (.swf).....	45
5.5 GEOGRAFISCHE INFORMATIE.....	46
5.5.1 Officiële standaarden.....	46
5.5.2 Defacto standaarden.....	46
5.5.2.1 Geography Markup Language (GML).....	46
5.5.2.2 GeoTIFF (.tif, .tiff).....	50
6. UITGEBREIDE INHOUDSOPGAVE.....	51