



STANDAARDEN VOOR BESTANDSFORMATEN

1. Woord vooraf

Belang

Om de lange termijn leesbaarheid van digitale documenten te verzekeren is het aangewezen om zoveel mogelijk gestandaardiseerde bestandsformaten te gebruiken. Standaarden spelen een belangrijke rol in de digitale bewaarstrategieën voor digitale archiefdocumenten zoals emulatie en migratie. Standaard bestandsformaten hebben hun status te danken aan (officiële) standaardiseringsinitiatieven of aan de grootschalige verspreiding.

Voordelen

Standaard bestandsformaten hebben het voordeel dat ze compatibel zijn met meerdere softwaretoepassingen. Wanneer één applicatie niet meer beschikbaar is, moeten de computerbestanden dus niet onmiddellijk omgezet worden. Een aantal conversiestappen worden bijgevolg vermeden. De interne bestandsstructuur van veel standaard bestandsformaten is publiekelijk verspreid zodat het in principe mogelijk is om ten allen tijde nieuwe viewers te maken. Het bewaren van informatie in bestandsformaten met de status van standaard dringt zich ook op om uitwisseling mogelijk te maken. Officiële standaarden of specificaties zijn niet afhankelijk van één producent, maar worden beheerd door standaardiseringsinitiatieven. Hierdoor wordt niet alleen een stukje stabiliteit gewaarborgd, de standaarden kunnen pas na het doorlopen van een hele procedure worden gewijzigd.

Parameters en opties

Bij de meeste bestandsformaten kunnen verschillende parameters en opties worden ingesteld. Vanuit archiveringsperspectief is het belangrijk om de nodige aandacht te besteden aan deze parameters en opties. Het technisch profiel en de mate waarin de digitale objecten geschikt zijn voor langetermijnarchivering worden immers sterk bepaald door de gehanteerde instellingen bij opslag. Volgende uitgangspunten zijn hierbij belangrijk en keren terug bij de meeste archiveringsformaten:

- hanteer de Unicode (evt. ASCII) tekenset bij het inbedden van metadata
- zorg ervoor dat de digitale objecten zo zelfvoorziening mogelijk zijn en zo weinig mogelijk afhankelijk van externe bronnen: alle elementen die nodig zijn voor een getrouwe reconstructie van het documenten worden maximaal aan de digitale objecten toegevoegd
- kies een ASCII-opslagmethode waar mogelijk: dit maakt uitwisseling tussen meerdere applicaties doorgaans gemakkelijker
- let op bij het toepassen van compressie: compressie kan resulteren in ongewenst gegevensverlies of tot softwareafhankelijkheid

Wrappers en codecs

Het hanteren van gestandaardiseerde bestandsformaten voor duurzame digitale archivering is belangrijk, maar is niet het enige aandachtspunt bij het kiezen van een

duurzaam opslagformaat voor gedigitaliseerde en digitale archiefdocumenten. Veel bestandsformaten zijn immers zogenaamde 'wrappers' of envelopformaten. Met andere woorden, deze formaten ondersteunen meerdere profielen en/of encodings en laten veel variaties toe. Zo ondersteunt het TIFF-formaat zowel LZW als JPEG compressie voor afbeeldingen. Een WAVE-bestand kan zijn geluidsdata in MP3-encoding bevatten. Een XML-document kan XML-dialect of een weinig semantisch documentmodel bevatten. Niet elk profiel of elke encoding is geschikt voor langetermijnarchivering zodat op voorhand goed moet worden afgewogen welk soort TIFF-, WAVE- of XML-bestanden worden samengesteld. Ook voor deze profielen en encodings worden bij voorkeur zo veel mogelijk standaarden gebruikt. In deze richtlijn wordt bij de wrapperformaten vermeld welke profielen of encodings aangewezen zijn.

Wanneer omzetten?

De digitale documenten kunnen op verschillende tijdstippen naar een standaard bestandsformaat worden omgezet.

1. Bewaar digitale documenten zoveel mogelijk van bij de creatie in het passende standaard bestandsformaat. Dit is vooral aangewezen bij audio-visuele documenten die je nog vele jaren nodig hebt. Bij tekstuele documenten is dit echter niet altijd mogelijk. Bijv. foto's → TIFF; audio → WAV
2. Zet de bestanden om wanneer hun onmiddellijk nut voorbij is. Bijv. MS WORD → PDF/A, ODF of TIFF
3. Zet de bestanden om terwijl de oorspronkelijke software nog operationeel is. Zorg eerst voor de omzetting, alvorens de software te vervangen. Bijv. AutoCAD → DXF of PDF/E
4. De omzetting dient ten laatste voor overdracht naar de archiefdienst te gebeuren.

2. Tekstdocumenten

FORMAAT	VERSIE	PARAMETERS en OPTIES	GEBRUIKEN VOOR
XML <i>eXtensible Markup Language</i>	1.0/1.1	<ul style="list-style-type: none"> – bij archiefdocumenten waarvan de lay-out belangrijk is: XSL-stylesheet opstellen – bij archiefdocumenten waarvan de structuur belangrijk is: XML Schema (of DTD) opstellen – baseer de granulariteit van het XML-document op de interne structuur van het archiefdocument en de gewenste zoekmogelijkheden – kies semantische XML-tags – zorg ervoor dat de XML-documenten leesbaar zijn voor meerdere applicaties – kies voor een XML-formaat dat vrij is van patentrechten die implementatie door derden in de weg staan of bemoeilijken 	<ul style="list-style-type: none"> – e-mails – tekstdocumenten – inhoud websites – databanken – metadata – spreadsheets
ODF <i>Open-Document Format</i>	1.0/1.2	<ul style="list-style-type: none"> – indien gewenst, vul metadata in de voorziene velden in of maak eigen metadatavelden aan. De metadata worden opgeslagen in het bestand ('meta.xml') 	<ul style="list-style-type: none"> – tekstdocumenten met grafieken en afbeeldingen – spreadsheets – presentaties
Unicode	5.2.0	<ul style="list-style-type: none"> – duidelijke veldscheidingstekens gebruiken – documenteer de velden in het Unicodebestand – documenteer de encoding en de linebreaks 	<ul style="list-style-type: none"> – databanken – logbestanden

PDF/A

Portable
Document Format
for Archiving

- gebaseerd op PDF 1.4
- moeten bevatten:
 - alle fonts
 - apparaatonaafhankelijke kleuren
 - XMP-metadata
 - ...
- mogen niet bevatten:
 - encryptie, beveiliging
 - LZW-compressie
 - ingebedde bestanden (audio, video, exe's)
 - transparantie
 - ...
- level A: bedoeld voor digital born documenten. Het volledige document is conform PDF/A (incl. tags en Unicode-mapping van alle karakters)
- level B: bedoeld voor gedigitaliseerde en digital born documenten. Bevat enkel die elementen die nodig zijn om de visuele verschijningsvorm te garanderen.

TIFF

6.0

Tagged Image
File Format

- scans: controleer of alles goed leesbaar is
veel gebruikte resolutie: 300 dpi
- bitdiepte:
 - zwart-wit: 1 bit
 - grijswaarden: 8 of 16 bits
 - kleur: 24 of 48 bits
- bij voorkeur compressieloos bewaren
- enkel baseline en vastgelegde TIFF-uitbreidingen toepassen: pas het TIFF-formaat niet intern aan
- RGB: afbeelding op scherm (sRGB: IEC 61966-2-1)/ CYMK: drukwerk
- byte volgorde (IBM of Macintosh): vrij te kiezen
- koppel metadata in de TIFF-header in:
 - 269: DocumentName
 - 270: ImageDescription
 - 271: Make
 - 272: Model
 - 305: Software
 - 306: DateTime
 - 315: Artist
 - 316: HostComputer
 - 33432: CopyRight

(X)HTML

HTML 4.01

(eXtensible)

HyperText

Markup Language

(ISO-15445)
of

XHTML 1.0

- (X)HTML-syntaxregels respecteren
- websites
- geen afgekeurde of niet-gestandaardiseerde tags en attributen
- nieuwsbrieven

- gebruiken
- leg bestandsgegevens (metadata) expliciet in de XHTML-header vast
(zie: *Digitaal Archiveren: Richtlijn 5. Websitesbeheer voor archivering.*)

3. Afbeeldingen

3.1 Rasterafbeeldingen

FORMAAT	VERSIE	PARAMETERS en OPTIES	GEBRUIKEN VOOR
TIFF <i>Tagged Image File Format</i>	6.0	<ul style="list-style-type: none"> – idem als bij tekstuele documenten – layers als afzonderlijke lagen in het TIFF-bestand bewaren – luchtfoto's: geografische referenties (coördinaten, projectie informatie, enz.) kunnen worden opgenomen in de geo-tags die GeoTIFF (versie 1.8.2, december 2000) voorziet. 	<ul style="list-style-type: none"> – digitale foto's in RAW-formaat – ingescande documenten – moederkopieën – luchtfoto's (in MrSid-formaat)
JPEG <i>Joint Pictures Experts Group</i>	JPEG	<ul style="list-style-type: none"> – bepaal de parameters in functie van de bestandsgrootte en de beeldkwaliteit – gebruik EXIF om metadata in te kapselen 	<ul style="list-style-type: none"> – digitale foto's in JPEG-formaat – foto's in websites – werk- of raadplegingskopieën
JPEG2000 <i>Joint Pictures Experts Group</i>	JPEG2000	<ul style="list-style-type: none"> – <i>lossless</i> of "visual lossless" compressie voor moederkopieën toepassen – <i>lossy</i> compressie voor werkkopieën toepassen – bewaar de afbeelding in een JPEG2000-bestand (.jp2), en niet als codestream (.jpc, .j2k). Gebruik de uitbreidingen op het JPEG2000-formaat (.jpx) enkel wanneer de afbeelding een kleurenprofiel nodig heeft dat niet in het gewone JPEG2000-formaat wordt ondersteund. – de afbeelding kan in meerdere resoluties in hetzelfde bestand worden opgeslagen – gebruik XML of XMP om metadata in te kapselen – gebruik GML om JPEG2000-afbeeldingen te georefereren (GMLJP2) 	<ul style="list-style-type: none"> – ingescande documenten – luchtfoto's (in MrSid-formaat) <p><i>opmerking: JPEG2000 wordt niet algemeen ondersteund door webbrowsers</i></p>
PNG <i>Portable Network Graphics</i>	1.0	<ul style="list-style-type: none"> – keuze tussen 8-bit en 24-bit 	<ul style="list-style-type: none"> – afbeeldingen (in websites) – afbeeldingen met paletkleuren (8 bit) of grijswaarden (24 bit) <p><i>opmerking: PNG past altijd compressie toe</i></p>
GIF	89	<ul style="list-style-type: none"> – beperking op aantal kleuren = 	<ul style="list-style-type: none"> – cartoons (met animatie)

- informatieverlies
- LET OP: Voor de creatie van GIF-bestanden dient men in principe te beschikken over een Unisyslicentie.
- logo's
- zwart-wit afbeeldingen
- afbeeldingen met paletkleuren en grote vlakken
- afbeeldingen in websites

3.2 Vectorafbeeldingen

FORMAAT	VERSIE	PARAMETERS en OPTIES	GEBRUIKEN VOOR
CGM <i>Computer Graphics Metafile</i>		<ul style="list-style-type: none"> – pas indien mogelijk een profiel toe – gebruik het WebCGM-profiel voor dynamische CGM-afbeeldingen met hyperlinks 	2 dimensionele vector afbeeldingen
DXF <i>Drawing eXchange Format</i>	25.1.01	<ul style="list-style-type: none"> – gebruik de ASCII-opslagmethode – in hoogst mogelijk versie bewaren 	2 en 3 dimensionele CAD/CAM-tekeningen
SVG <i>Scalable Vector Graphics</i>	1.2	<ul style="list-style-type: none"> – kan animatie en interactieve elementen bevatten – metadata kan in XML-vorm aan de afbeeldingen worden toegevoegd 	<ul style="list-style-type: none"> – 2 dimensionele afbeeldingen – flowcharts – animaties
PDF/E <i>Engineering document format using PDF</i>	1	<ul style="list-style-type: none"> – gebaseerd op PDF versie 1.6 – moeten bevatten: <ul style="list-style-type: none"> ▪ fonts ▪ apparaatonafhankelijke kleuren ▪ XMP-metadata 	2 dimensionele CAD/CAM tekeningen

4. Geluid

FORMAAT	VERSIE	PARAMETERS en OPTIES	GEBRUIKEN VOOR
PCM <i>Pulse Code Modulation</i>		<ul style="list-style-type: none"> – sample-rate en bitdiepte kiezen in functie van de kwaliteit van de analoge bron – DAT-bestand bijleveren waarin de sample-rate en sample-resolutie expliciet zijn vastgelegd 	digitale moederkopieën
WAVE <i>Waveform Audio File Format</i>		<ul style="list-style-type: none"> – sample-rate en bitdiepte kiezen in functie van de kwaliteit van de analoge bron. CD-kwaliteit is 44,1 KHz en 16 bit/sample. – codecs: <ul style="list-style-type: none"> – master/moederkopie: lineaire pulse code modulation – uitwisseling/streaming: MP3-codec – gebruik 'WAVE_FORMAT_EXTENSIBLE' wanneer: <ul style="list-style-type: none"> – een hogere bitdiepte dan 16 	digitale moederkopieën of uitwisseling

		bits/sample nodig is	
		<ul style="list-style-type: none"> - het aantal kanalen hoger is dan 2: <i>Wave Format Extensible</i> ondersteunt max. 18 kanalen. - de mapping van kanalen naar speakers moet worden vastgelegd - gebruik BWF/RF64¹ wanneer: <ul style="list-style-type: none"> - de bestandsomvang groter is dan 4 GB. - het aantal kanalen hoger is dan 2: zie <i>Wave Format Extensible</i> 	
MP3	1.0	<ul style="list-style-type: none"> - bepaal de parameters frequentie, bitrate en aantal kanalen in functie van de gewenste kwaliteit en bestands grootte 	uitwisseling van geluidsbestanden via netwerken. Gebruik van streaming MP3 is mogelijk.

5. Beeld en geluid

FORMAAT	VERSIE	PARAMETERS en OPTIES	GEBRUIKEN VOOR
MXF <i>Material Exchange Format</i>	1.0	<ul style="list-style-type: none"> - MXF heeft diverse 'operational patterns'. MXF OP1A is de aangewezen pattern om de beeld- en geluids informatie samen te bewaren (SMPTE 378M-2004: Operational Pattern 1a (Single Item, Single Package)). - MXF is een zgn. wrapperformaat en kan beeld- en geluids informatie (de 'essence') in diverse encodings bevatten. Volgende encodings zijn aangewezen voor: <ul style="list-style-type: none"> ▪ videobanden: lossless Motion JPEG-2000 ▪ analoge masters: <ul style="list-style-type: none"> - beeld: uncompressed YUV 4:2:2 met hoge bitdiepte (8 of 10) - geluid: uncompressed PCM (48 KHz en 24 bits bitdiepte) ▪ digitale masters: <ul style="list-style-type: none"> - video: uncompressed YCbCr, 4:2:2 of 4:4:4 met hoge bitdiepte (8 of 10) - geluid: uncompressed PCM (48 KHz en 24 bits bitdiepte) 	hoge resolutie digitale moederkopieën (essence + metadata) die nog editeerbaar moeten zijn

¹ De European BroadCasting Union legde het Broadcast Wave Format (BWF) als uitwisselingsformaat voor audio vast. BWF is een verfijning van het algemene WAVE-formaat. BWF voorziet een extra 'chunk' voor de opslag van specifieke metadata (chunknaam 'bext'). BWF ondersteunt enkel lineaire PCM en MPEG als audiocodec. Het standaard BWF-profiel is lineaire PCM, 16 bits/sample en 48 KHz. RF64 biedt de mogelijkheid om BWF-bestanden groter dan 4 GB en met multichanneling op te slaan.

		– Inkapselen van beschrijvende en technische metadata in XML	
MJPEG2000 <i>Motion JPEG2000</i>	1.0	<ul style="list-style-type: none"> – MJPEG2000 Simple Profile: <ul style="list-style-type: none"> ▪ 1 videotrack: max. 30 frames/seconde ▪ max 1 geluidtrack: max. 48Khz, 8 of 16 bits, 1 of 2 kanalen, PCM-codec ▪ geen externe bronnen of referenties: alle gegevens zijn in het bestand opgeslagen – opslag: in een MXF-wrapper, of rechtstreeks in MJPEG2000 <ul style="list-style-type: none"> ▪ als 1 digitaal object die alle metadata, video- en audio informatie bevat ▪ als meerdere digitale objecten: het MJPEG2000 metadata-bestand verwijst naar de afbeeldingen (bijv. JPEG2000) en het audiobestand (bijv. WAV). – compressie: <i>lossless</i> of <i>visual lossless</i> – voeg metadata in XML-vorm toe 	hoge resolutie digitale moederkopieën die niet of slechts uitzonderlijk editeerbaar hoeven te zijn
MPEG <i>Moving Pictures Experts Group</i>	1.0	bepaal het aantal frames (bijv. 30), de kleurdiepte (8, 16, 24) en de resolutie (origineel, 160 x 120, 320 x 240, 640 x 480) in functie van de gewenste kwaliteit en bestandsgrootte	distributie van bewegende computerbeelden met/zonder geluid (VHS-kwaliteit)
MPEG (* .m2v)	2.0	idem als bij MPEG 1.0	digitale televisie, multimedia (DVD-kwaliteit).

6. ? Vragen ? Suggesties ?



Met al uw vragen en suggesties kan u terecht bij eDAVID: <mailto:info@edavid.be>



Meer info: F. Boudrez, *Standaarden voor digitale archiefdocumenten*, (<http://www.edavid.be/davidproject>).