Nationaal Archief
Ministerie van Onderwijs, Cultuur en Wetenschap

Bestandsstrategieën Nationaal Archief

VERSIE 1.0

Datum 15-11-2016
Status Definitief

Colofon

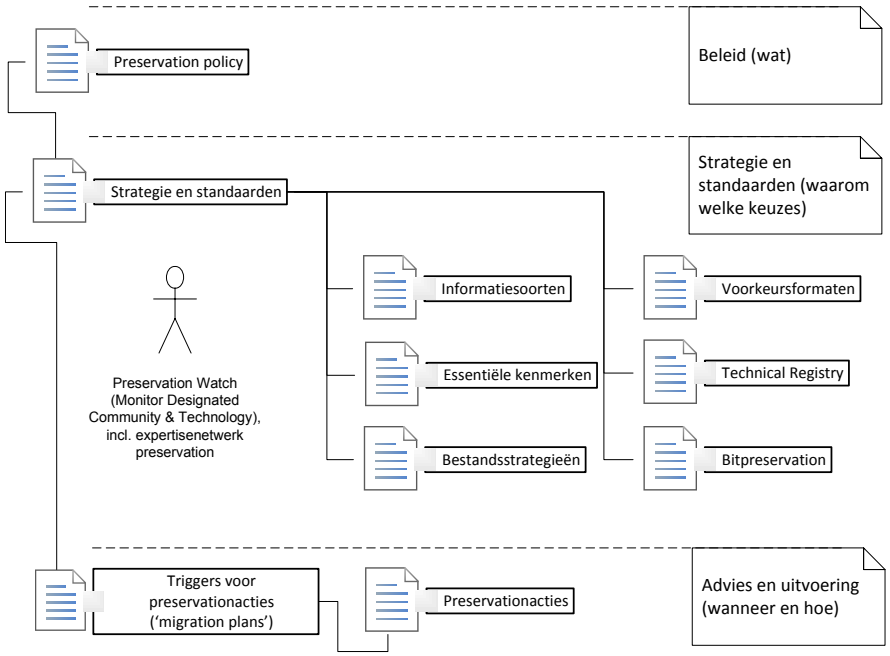
Projectnaam	Ontwikkelen duurzaamheidsstrategieën
Projectleider(s)	Remco van Veenendaal
Contactpersoon	R. van Veenendaal T +31 6 29 45 19 51 F +31-70-331 5477 remco.van.veenendaal@nationaalarchief.nl Postbus 90520 | 2509 LM Den Haag
Auteurs	R. van Veenendaal, Pepijn Lucker
Versie	1.0
Bijlage(n)	

Inhoud

1 Inleiding

Vanuit de (rijks)overheid verwacht het NA de komende jaren vooral digitale informatieobjecten[1] – veelal digital born – in bepaalde bestandsformaten te ontvangen: TIFF-scans, Outlook e-mails, PDF-documenten en Microsoft Office-bestanden. Omdat het NA een kwaliteitsniveau boven bitstreampreservation – het in stand houden van de bitstreams van de digitale informatieobjecten – voorstaat, ontwikkelen we (preservation)strategieën voor het omgaan met die bestandsformaten.

De strategieën voor het omgaan met deze bestandsformaten zijn verzameld in dit document, voorafgegaan door de algemene uitgangspunten van de preservationfunctie van het NA. Dit document vormt samen met andere componenten[2] de uitwerking van de preservation policy, waarin het overkoepelende preservation beleid van het NA is beschreven (zie afbeelding).



Figuur 1 - overzicht componenten preservation NA

Dit is een levend document: periodiek zal de lijst bestandsformaten uitgebreid worden en zullen de beschreven bestandsstrategieën worden bijgewerkt. Hierbij kijken we vooral naar de importantie (bijv. om hoeveel bestanden gaat het) en urgentie (bijv. wanneer is de strategie nodig) van het toevoegen van een bestandsformaat.

[1] Informatie object wordt in het OAIS model omschreven als "A Data Object together with its Representation Information". We volgen in dit document het "REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)" (Magenta Book, 2012).

[2] NB: Deze documenten hangen met elkaar samen maar kunnen ook elk afzonderlijk van elkaar worden gelezen.

2 Doel en resultaat

2.1 Doel

Het NA wil ook in de toekomst authentieke en betrouwbare informatie beschikbaar kunnen stellen. Het NA staat zoals gezegd een kwaliteitsniveau boven de bitstreampreservation voor en wil op basis van informatie over de vorm, inhoud en structuur van informatieobjecten beter kunnen anticiperen op veranderingen in techniek en de gevolgen van die veranderingen, en tijdig ingrijpen indien de authenticiteit en toegankelijkheid van informatieobjecten bedreigd wordt.

2.2 Doelgroep

Dit document is in eerste instantie bedoeld voor informatieprofessionals bij het NA die werken aan het duurzaam toegankelijk houden van het bestaande digitale archief van (met name) de rijksoverheid. Maar ook de RHC's, die werken met dezelfde digitale infrastructuur van het NA, kunnen van dit document gebruik maken.

2.3 Resultaat

Om succesvol te kunnen anticiperen op veranderingen m.b.t. bestandsformaten van informatieobjecten, ontwikkelt het NA bestandsstrategieën voor bestandsformaten. Vanuit de (rijks)overheid verwacht het NA de komende jaren vooral TIFF-scans, Outlook e-mails, PDF-documenten en Microsoft Office-bestanden te ontvangen. In deze versie van dit document staan daarom de bestandsstrategieën voor de volgende bestandsformaten:

- TIFF (.tif/tiff)
- E-mail (o.a. .msg, .eml)
- PDF en PDF/A (.pdf)
- Microsoft Office Word (.doc, docx)
- Microsoft Office Excel (.xls, .xlsx)
- Microsoft Office Powerpoint (.ppt, .pptx)
- Microsoft Office Access (.mdb, .accdb)

Deze lijst zal in volgende versies van dit document worden gereviewed en kan dan worden uitgebreid met bijvoorbeeld strategieën voor bestandsformaten voor websites, social media of databases (anders dan MS Access).

3 Bestandsstrategieën

3.1 Algemene uitgangspunten

In deze sectie staat beschreven wat momenteel de algemene uitgangspunten zijn voor preservation bij het NA. In principe gelden die dus voor iedere specifieke bestandsstrategie die in dit document beschreven wordt. Periodiek worden de eerdergenoemde preservation documenten geëvalueerd en bijvoorbeeld vanwege voortschrijdend inzicht aangevuld en/of aangepast.

De Rijksoverheid stimuleert het gebruik van open data, open standaarden en opensourcesoftware. De Nederlandse overheid hanteert daarbij het principe *pas toe of leg uit.* De Archiefregeling 2009 stelt dat digitale informatie "*uiterlijk op het tijdstip van overbrenging, [is] opgeslagen in een valideerbaar en volledig gedocumenteerd bestandsformaat dat voldoet aan een open standaard."*

Open standaarden worden door het Forum Standaardisatie (de organisatie die de interoperabiliteit en de toepassing van open standaarden binnen de Nederlandse overheid bevordert) omschreven als:

"*'Open' heeft betrekking op het standaardisatieproces en is één van de toetsingscriteria voor opname op de lijsten. Het gaat daarbij om laagdrempelige beschikbaarheid van documentatie, geen hindernissen op basis van intellectuele eigendomsrechten (bijv. geen patent royalties), inspraakmogelijkheden, en onafhankelijkheid en duurzaamheid van de standaardisatie-organisatie."*[3]

"*Overheden en semi-overheden zijn verplicht de open standaarden die op de lijst staan, bij aanschaf of (ver)bouw van ICT-systemen/-diensten te eisen ('pas toe'). Afwijken mag alleen met zwaarwegende redenen en verantwoording hierover moet worden afgelegd in het jaarverslag ('leg uit'). 'Pas toe of leg uit'-standaarden zijn open standaarden waarvoor breed draagvlak bestaat maar die nog niet breed geadopteerd zijn. Daarom krijgen deze standaarden de status van 'pas toe of leg uit'."*[4]

Indien vlak voor overbrenging informatie moet worden omgezet naar een open standaard/formaat is het raadzaam om hierover vooraf advies te vragen aan het NA aangezien er bij deze omzetting ongewenst informatieverlies kan optreden. Het NA kan dan adviseren over de technische kanten en de essentiële kenmerken van informatieobjecten. Dit document behandelt vooral de technische kanten van bestandsformaten van informatieobjecten. Het refereert niet aan archivistische keuzes waar je tijdens het uitvoeren van een strategie voor komt te staan met betrekking tot essentiële kenmerken (structuur, vorm, inhoud, gedrag en context) van informatieobjecten. Het is belangrijk om je realiseren dat alleen een technische benadering wel nodig, maar niet voldoende is voor optimale preservation.

Het NA beschrijft in een apart document hoe wordt omgegaan met essentiële kenmerken. Dit document kan worden gebruikt om samen met de archiefvormer de essentiële kenmerken van informatieobjecten te kunnen vaststellen en de mate waarin die door de tijd heen worden gepreserveerd. Ook bij eventuele latere omzettingen uitgevoerd door het NA houden we rekening met de essentiële kenmerken.

Het NA heeft ook een lijst van voorkeursformaten opgesteld die preservation en daarmee duurzame toegankelijkheid van informatie makkelijker maakt.

[3] Bron: https://www.forumstandaardisatie.nl/open-standaarden/over-open-standaarden/ (geraadpleegd 27-11-2015)

[4] Bron: https://www.forumstandaardisatie.nl/open-standaarden/lijsten-met-open-standaarden/ (geraadpleegd 19-11-2015)

Het NA gaat uit van custodial preservation, oftewel de preservation van informatie die in het e-Depot van het NA is opgeslagen. We sluiten niet uit dat we op termijn aan post-custodial[5] en dus meer gedistribueerde preservation gaan doen.

Het NA kiest ook voor just-in-time-preservation: het ingrijpen wanneer het nodig is, dus als veranderingen de goede, geordende en toegankelijke staat van informatieobjecten bedreigen. We kiezen hiermee niet voor just-in-case-preservation, wat inhoudt dat ingegrepen wordt zodra het kan. In een just-in-time-scenario grijpen we tijdig en met de meest actuele kennis en middelen in.

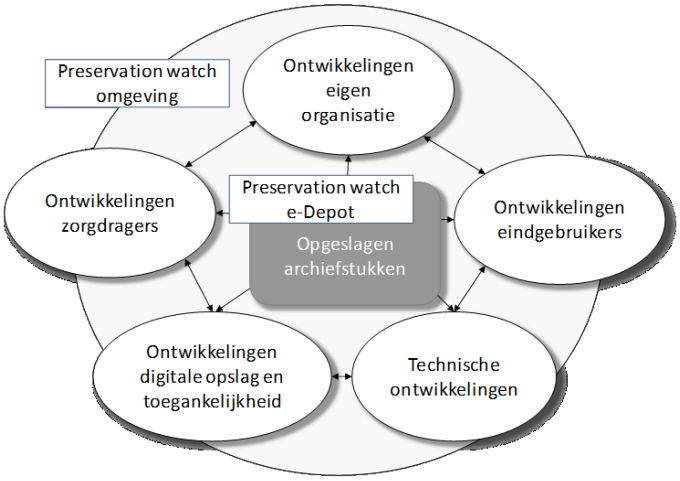
Aan de basis van de preservationfunctionaliteit ligt bitpreservation. Dat wil zeggen dat er een bitstream is die te allen tijde een bit-perfect copy oplevert. Het NA gaat voor deze bit-perfect copy maar zorgt er tevens voor dat informatie in de toekomst authentiek en betrouwbaar beschikbaar gesteld wordt en dat is meer dan bitstream preservation. De e-Depot-functionaliteit is ingericht voor bitpreservation door:
- het onderhouden van op zijn minst een beschikbare kopie van elke bitstream. Het NA slaat dus altijd minstens twee manifestaties[6] van iedere bitstream op: het origineel en minstens een kopie;
- het garanderen van de integriteit van de bitstream (checksum controleren en cyclus van controle instellen);
- het aantonen en documenteren van bovenstaande.

Het NA maakt zoveel mogelijk gebruik van automatisering bij opname van informatieobjecten, het uitvoeren van beheer, preservationacties en acties met betrekking tot de toegankelijkheid. We beschikken hiervoor over een e-Depot, dat via preservation workflows migratie als voornaamste strategie kent, waarmee we bijvoorbeeld emulatie als alternatieve strategie niet uitsluiten.

Het NA heeft een preservation watch ingesteld. Deze watch dient om de reikwijdte te bepalen van de technologieën om de informatieobjecten en metadata te managen en hier toegang tot te verlenen, de support van die technologieën in de organisatie en community te monitoren en triggers in te bouwen. Praktisch betekent dat:
- Het bijhouden van de (inter)nationale ontwikkelingen op het gebied van technologische veranderingen en standaarden en gebruikte hard- en software door producers en daar rapport over uit brengen;
- Het regelmatig herzien van designated communities;
- Het uitvoeren van risico-inventarisaties op de informatieobjecten en metadata in de e-Depot-voorziening.
- Het monitoren van de Producer, Consumer en interne organisatie op veranderingen die invloed kunnen hebben op de duurzame toegankelijkheid van de informatieobjecten.

[5] Post-custodial preservation houdt voor ons in dat de te preserveren informatieobjecten (waarvoor we als NA verantwoordelijk zijn) zich niet bij het NA bevinden.

[6] Met manifestaties wordt binnen de context van het e_Depot bedoeld dat van een informatieobject meerdere representaties worden bewaard, bijv. een MS Word manifestatie en een PDF manifestatie. Of een TIFF en een JPEG manifestatie.



Indien een preservationactie noodzakelijk is, dient op basis van actuele bestandsstrategieën een preservationplan gemaakt te worden. De verplichte onderdelen daarvan zijn:

- Een definitie van het type informatieobject waar het op van toepassing is;
- Een beschrijving van de verandering;
- Een beschrijving van de beoogde uitkomst;
- Een stappenplan (incl. naam en versie van de te gebruiken soft- en hardware, noodzakelijk vereiste configuraties, en de exacte volgorde van de benodigde stappen);
- Succesfactoren;
- Testen, goedkeuren en documenteren van het proces.

3.2 Opbouw hoofdstukken

Na de inleiding in hoofdstuk1, uitleg over het doel en het resultaat van dit document in hoofdstuk 2 en het huidige hoofdstuk met algemene uitgangspunten en informatie, volgen hoofdstukken met strategieën voor specifieke bestandsformaten. Deze inhoudelijke hoofdstukken hebben steeds dezelfde opbouw. In paragraaf 1 geven we algemene informatie over het bestandsformaat, in paragraaf 2 volgt een opsomming van de (belangrijkste) risico's en in paragraaf 3 evalueren we op hoofdlijnen het bestandsformaat. Hoe het e-Depot met een bestandsformaat kan omgaan staat in paragraaf 4. Deze informatie verdelen we aan de hand van de informatie uit de Registry van het e-Depot[7]: informatie over bestandsformaten, over migratiepaden en over software en tools. Voor de meeste bestandsformaten zijn alternatieven mogelijk en de belangrijkste staan in paragraaf 5. Tenzij anders aangegeven beperken we de lijst alternatieven tot die bestandsformaten die we voldoende geschikt vinden voor duurzame toegankelijkheid. In paragraaf 6 sluiten we steeds af met een voorgestelde strategie.

[7] Preservica heeft een Registry met bestandsformaten (File Formats), migratiepaden (Migration Pathways), software (Software) en tools (Tools). De Registry heef ook informatie over informatiesoorten (Property Groups) en daarover staat meer in paragraaf **Fout! Verwijzingsbron niet gevonden.**.

4 TIFF

4.1 Algemene informatie

Volgens Wikipedia (https://en.wikipedia.org/wiki/Adobe_TIFF) is het Tagged Image File Format (TIFF) een formaat voor rasterafbeeldingen. Het is ontwikkeld door de Aldus Corporation. TIFF is primair bedoeld voor scannen en desktop publishing. In 1994 nam Adobe Systems Incorporated Aldus over, en kreeg daarmee de rechten op het TIFF-formaat. Adobe heeft TIFF sindsdien onderhouden.

De meest recente versie van TIFF is 6.0, uit 1992. Vanwege de achterwaartse compatibiliteit zijn eerdere versies van TIFF ook geldige TIFF 6.0-versies. (Deze compatibiliteit zit overigens vooral in deel 1 van het bestandsformaat, de Baseline. In deel 2, de Extensions, kunnen leverancierspecifieke extensies zitten die de compatibiliteit negatief beïnvloeden.) De meest gangbare en gebruikte TIFF-versies zijn 5 en 6.

PRONOM (http://apps.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1099) meldt dat TIFF-bestanden uit drie delen bestaan: een Image File Header (IFH), een Image File Directory (IFD), en de afbeeldingsdata. TIFF-bestanden kunnen meerdere afbeeldingen bevatten (multipage TIFF) en iedere afbeelding heeft een eigen IFD. De IFH staat altijd aan het begin van het bestand en wordt gevolg door een pointer naar de eerste IFD. De IFD bevat metadata die de bijbehorende afbeelding beschrijft, opgeslagen als een lijst tags. De IFD bevat ook een pointer naar de eigenlijke afbeeldingsdata.

TIFF ondersteunt kleurdieptes (colour depth) van 1 tot 24 bit (monochrome tot true colour), ongecomprimeerde data en een waaier aan compressietypes (RLE, LZW, CCITT Group 3 en Group 4, en JPEG).

Meer informatie over TIFF:
- https://en.wikipedia.org/wiki/Adobe_TIFF (algemene informatie over TIFF)
- http://wiki.dpconline.org/images/f/f3/TIFF_Assessment_v1.2_external.pdf (preservationassessment van TIFF door het British Library Preservation Team)
- http://www.archives.gov/preservation/products/definitions/tif.html (preservationinformatie van de National Archives in de V.S. m.b.t. TIFF)
- http://www.digitalpreservation.gov/formats/fdd/fdd000022.shtml (preservationinformatie van de Library of Congress in de V.S. m.b.t. TIFF)
- http://www.dlib.org/dlib/july08/buonora/07buonora.html (over de trend om TIFF te vervangen door JPEG 2000)
- http://www.aquaforest.com/en/tiff_versus_pdf.asp (over PDF als alternatief voor TIFF)
- http://apps.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=1099 (PRONOM over TIFF)

Een interessante, maar nog jonge, ontwikkeling is het initiatief om, á la PDF/A en parallel aan de ontwikkeling van de DPF Manager, een validatietool voor TIFF, een TIFF/A ISO-standaard te ontwikkelen, zie http://tiff-a.org/. Het verdient aanbeveling deze ontwikkeling in de gaten te houden.

4.2 Risico-inventarisatie

Voordelen van het TIFF bestandsformaat:

- TIFF is in staat bi-level, grijswaarden, pallette-color en full-color afbeeldingen te beschrijven in verschillende kleurruimten.
- Het formaat is niet gerelateerd aan specifieke camera's, scannners, printers of soft- en hardware.
- Het formaat is uitwisselbaar, het geeft geen voorkeur aan bepaalde besturingssystemen, compilers of processoren.
- TIFF kan zich verder ontwikkelen wanneer zich nieuwe eisen voordoen, zonder dat de baseline TIFF hierdoor verandert.
- De eenvoud en de openheid van het formaat

Nadelen van het TIFF bestandsformaat zijn:

- Er is geen standaard manier voor het aangeven van multi-layer relaties voor een bestand wat uit meerdere TIFF pagina's bestaat.
- TIFF heeft een beperking in bestandsgrootte van 4 GB.
- TIFF staat toe dat er informatie door derden aan het bestand kan worden toegevoegd. Dit kan ongedocumenteerde of “encrypted” informatie zijn.
- TIFF is eigendom van Adobe. Dit bedrijf is ook eigenaar van ingewikkeldere formaten zoals het Adobe PhotoShopDocument (PSD) formaat waaraan de laatste 10 jaar meer aandacht is besteed mbt de ontwikkeling van deze formaten. De ontwikkeling van het TIFF formaat heeft daardoor stil gestaan.

4.2.1 Extensies

Bij de TIFF-bestanden kan het voorkomen dat er binnen de bestanden bepaalde extensies (uit deel 2) worden gebruikt die specifiek zijn voor een bepaalde fabrikant. Deze extensies zijn vaak slecht of niet gedocumenteerd. Dit komt voornamelijk voor bij TIFF-bestanden die rechtstreeks met een digitale camera gemaakt zijn. De camerafabrikant voegt in daarvoor bestemde tags bedrijfseigen informatie toe. Nikon staat er bijvoorbeeld om bekend deze informatie te voorzien van encryptie.

4.2.2 Specifieke kleurruimtes

Het gebruik van specifieke kleurruimtes die niet in de baselinespecificaties van TIFF worden beschreven kan in de toekomst problemen opleveren.

4.2.3 Softwareondersteuning multipage-TIFF-bestanden

Het eerder genoemde British Library Preservation Team ontdekte dat software soms bij multipage-TIFF-bestanden (baseline 6.0) alleen de eerste afbeelding toont.

4.2.4 Beperkte detectie van corruptie of beschadiging

Ook is de ondersteuning van tools voor het detecteren van corrupte of beschadigde TIFF-bestanden beperkt. Bijv. JHOVE kan soms onterecht een beschadigd TIFF-bestand als valide en welgevormd bestempelen.

4.2.5 Ter info: black pixel detector

Specifiek voor het analyseren van scans in TIFF-formaat is in 2012 door het NA (Maurice de Rooij) een "black pixel detector" ontwikkeld[8]. Deze black pixel detector analyseert TIFF-bestanden en laat (in Microsoft Office Excel) zien of een bestand puur zwarte of witte pixels bevat. In theorie mogen deze in scans niet voorkomen, dus zijn ze een goede indicatie voor beschadigingen.

4.2.6 LZW-compressie

Meerdere bronnen geven aan dat TIFF diverse vormen van (lossy en lossless) compressie toestaat, waaronder JPEG. In het verleden werd soms voor (lossy) JPEG-compressie gekozen, bijv. omdat (lossless) LZW-compressie gepatenteerd was en andere compressiemethoden onvoldoende ondersteund werden. LZW-patenten zijn echter sinds 2004 verlopen en vanuit dat perspectief bekeken kan tegenwoordig veilig gekozen worden voor TIFF-bestanden met lossless LZW-compressie.

4.3 Evaluatie

TIFF is een eenvoudig, apparatuuronafhankelijk, besturingssysteemonafhankelijk, goed uitwisselbaar, 'industry standard' bestandsformaat met een achterwaarts compatibele stabiele kern (TIFF-baseline).

Bij nieuwe TIFF-versies moet vooral onderzocht worden in hoeverre derden in de extensies (ongedocumenteerde of ge-encrypte) informatie toevoegen, waardoor die versies in de toekomst mogelijk onleesbaar worden.

4.4 Ondersteuning in het e-Depot

4.4.1 Formaten

Negen formaten voldoen aan de zoekterm "tagged" (waarvan vier ook worden gevonden via de zoekterm "tiff"). De hoofdingang voor het TIFF-formaat is PRONOM Unique Identifier (PUID[9]) fmt/353. De andere (versiespecifieke) PUIDs zijn verouderd (deprecated).

PUID	Name	Version
fmt/155	Geographic Tagged Image File Format (GeoTIFF)	
fmt/10	Tagged Image File Format	6
fmt/353	Tagged Image File Format	
fmt/7	Tagged Image File Format	3
fmt/8	Tagged Image File Format	4
fmt/9	Tagged Image File Format	5
fmt/154	Tagged Image File Format for Electronic Photography (TIFF/EP)	
fmt/153	Tagged Image File Format for Image Technology (TIFF/IT)	
fmt/156	Tagged Image File Format for Internet Fax (TIFF-FX)	

[8] Zie S:\Digitale Innovatie\Infrastructuur\Preservation\R&D\afbeeldingen\black_pixel_detector\, geraadpleegd op 10-08-2015.

[9] Zie http://www.nationalarchives.gov.uk/PRONOM/Default.aspx, geraadpleegd op 10-08-2015.

4.4.2 Migration Pathways

Het e-Depot kent tientallen migration pathways om formaten van en naar TIFF om te zetten, incl. omzetting naar PNG, JPEG, PDF, PDF/A, GIF en JPEG2000. Dit hoge aantal reflecteert de mate waarin TIFF ondersteund wordt.

4.4.3 Software en Tools

Beschikbare software/tools voor het omgaan met TIFF zijn bijv. Java ImageIO en Jhove-TIFF. Ook de standaard e-Depotcomponenten identifyFileComponents en identifyWebpages (her)kennen TIFF.

PUID	Name	Version
tool/102974	Java ImageIO GIF to TIFF	
tool/103555	Java ImageIO JPEG 2000 to TIFF	
tool/105015	Java ImageIO JPEG to TIFF	
tool/105504	Java ImageIO PNG to TIFF	
tool/105562	Java ImageIO TIFF to GIF	
tool/105620	Java ImageIO TIFF to JPEG	
tool/105678	Java ImageIO TIFF to JPEG 2000	
tool/105748	Java ImageIO TIFF to PNG	
tool/106233	Jhove-Tiff	

4.5 Alternatieven

Een alternatief voor TIFF is PDF en met name PDF/A voor archivering en preservering. PDF wordt elders in dit document besproken.

Een reden waarom PDF een alternatief is voor TIFF is eerder genoemd: Adobe heeft de afgelopen jaren weinig geïnvesteerd in het TIFF-formaat, en bijv. meer in PDF. Daarnaast kan TIFF standaard geen (gescande) tekst opslaan, terwijl PDF daar wel mogelijkheden voor heeft. PDF heeft (standaard) ook uitgebreidere mogelijkheden voor metadatering (bijv. via XMP[10]) en documentstructuur. TIFF heeft standaard alleen mogelijkheden voor opeenvolgende pagina's, PDF bijv. ook voor Bookmarks, Hyperlinks, Tags en Annotaties.

Hoewel TIFF nog altijd een goed formaat is voor het opslaan van (scans van) plaatjes, lijkt PDF hier steeds vaker voor gebruikt te (gaan) worden, vooral omdat dit formaat actief wordt doorontwikkeld en meer functionaliteiten biedt dan het TIFF-formaat. Zie ook http://www.aquaforest.com/en/tiff_versus_pdf.asp.

Een ander alternatief voor TIFF is JPEG 2000 (.jp2, .jpx). *"JPEG2000 is de meest recente JPEG-standaard die (in 2000) is ontwikkeld door het JPEG-comité (Joint Photographic Experts Group). De eerste standaard die (in 1992) door dit comité werd ontwikkeld is de JPEG-standaard. Deze zogenaamde baseline JPEG-compressie is een lossy compressie. Dit betekent dat wanneer de compressie is toegepast het niet meer mogelijk is het originele ongecomprimeerde beeld (zoals bijvoorbeeld een scan) uit het bestand te decoderen. Bij een hoge compressie kan het dataverlies in*

[10] Zie bijv. http://www.adobe.com/products/xmp.html, geraadpleegd op 11-08-2015.

eerste instantie met het blote oog niet waarneembaar zijn, maar het is wel aanwezig.
De nieuwe JPG2000 standaard is door het comité ontwikkeld om te voldoen aan de volgende eisen:
- *een compleet open formaat*
- *een verbeterend lossy compressie-algoritme in vergelijking met de huidige JPEG-standaard (let op: het NA adviseert geen lossy compressie te gebruiken[11])*
- *een optie voor lossless compressie (wat, als er compressie gebruikt wordt, beter gebruikt kan worden dan lossy compressie)*
- *volwaardige metadata-ondersteuning*
- *een progressieve weergave van bestanden (mogelijkheid tot verschillende resoluties en verschillende schalen)*
- *de mogelijkheid tot het verwerken van grote bestanden met een hoge dynamische weergave*

JPEG 2000 is een uitgebreide standaard die bestaat uit verschillende gedeelten (parts) en gebruik maakt van de waveletcompressiemethode. Het bestandsformaat bestaat niet alleen uit compressie-algoritmen maar biedt ook ondersteuning voor het gebruik van de standaard in verschillende applicaties. De voor het NA interessante (binaire) toepassingen van de standaard zijn:
- *JP2 is het basisformaat dat enkel gebruikt wordt voor afbeeldingen met een beperkte set voor kleurencodering (.jp2)*
- *JPX is een uitgebreide versie van de JP2-standaard. Er is ondersteuning voor meerdere afbeeldingen en er zijn meer mogelijkheden voor kleurencoderingen (.jpx)*
- *MJ2 (MotionJPEG2000) is de standaard voor bewegend beeld, waarbij elk frame een gecomprimeerd JPEG2000-bestand is (.mj2)*
- *JPM is een formaat voor samengestelde bestanden die kunnen bestaan uit afbeeldingen, grafische bestanden en tekst (.jpm)*

JPEG2000 is een standaard die met één compressie-algoritme zowel een lossless (reversible) als een lossy (irreverisble) compressie kan uitvoeren. Het JP2-formaat definieert de kern-decoder en focust zich op de codestream van het bestand. Dit is de collectie van bits die de gecomprimeerde data bevatten en de parameters die nodig zijn om deze data te kunnen interpreteren (decoderen)." (vrije vertaling en samenvatting van http://www.dpconline.org/component/docman/doc_download/87-jpeg-2000-a-practical-digital-preservation-standard)

Hoewel JPEG 2000 als ISO-standaard is gepubliceerd, wordt ze (bijvoorbeeld in webbrowsers) nog niet volop (native) ondersteund. Zie bijvoorbeeld https://en.wikipedia.org/wiki/JPEG_2000 voor meer algemene informatie en http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/ voor een blog over de (mogelijke) risico's van het gebruik van JPEG 2000.

Op termijn kunnen opensourceimplementaties van de JPEG 2000-specificatie interessant zijn om nader te verkennen. Zie bijvoorbeeld https://en.wikipedia.org/wiki/OpenJPEG,

[11] Op het NA-intranet staat hier "sanctioneerd" (sic.), wat (op het moment van schrijven, in 2015) achterhaald is omdat het NA bijv. geen (substitutie)machtigingen meer verleent, maar een adviserende rol heeft.

https://github.com/uclouvain/openjpeg en
https://en.wikipedia.org/wiki/JasPer.

4.6 Voorgestelde strategie

TIFF was, is en blijft voorlopig een goed bestandsformaat voor opslag en beschikbaarstelling van beeldmateriaal, en in het bijzonder scans. Als vanwege het verminderen van de bestandsgrootte voor compressie gekozen wordt, dan kan beter een lossless methode worden gekozen dan een lossy. Risico's m.b.t. het omgaan met TIFF-bestanden gaan vooral over ondersteuning van software/tools voor specifieke extensions.

De strategie voor het omgaan met TIFF is dat tot nader order, mede vanwege onze just-in-time-strategie, geen preservationacties nodig zijn. Periodiek, bijv. jaarlijks, moet gecontroleerd worden of oude TIFF-versies nog ondersteund worden en of nieuwe TIFF-versies en/of TIFF-software/tools ontwikkeld zijn. Hierbij moeten ook de ontwikkelingen m.b.t. de alternatieven voor TIFF meegenomen worden. Deze controleslagen kunnen triggers zijn voor het aanpassen van de strategie.

NB: TIFF komt niet voor op de lijst van Open Standaarden van het Forum Standaardisatie.

5 E-mail

5.1 Algemene informatie

Het begrip e-mail dekt twee ladingen: enerzijds het e-mailsysteem dat langs elektronische weg berichten transporteert en anderzijds de e-mailberichten zelf. Een e-mailsysteem bestaat uit programmatuur, transportmedium, (zoals netwerkvoorzieningen) en computers. Het stelt mensen in staat asynchroon berichten uit te wisselen van en naar elektronische postbussen. We bespreken het e-mailsysteem hier slechts kort en focussen ons vooral op de (preservering van de) berichten.

Een e-mailbericht wordt door (eind)gebruikers meestal ervaren als één geheel, maar bestaat eigenlijk uit twee delen: een kop (message header) en de inhoud van het bericht (message body). Een message header bevat informatie over het bericht, zoals afzender, geadresseerde, onderwerp, datum en andere gegevens. Het bericht zelf, inclusief eventuele bijlagen, kan data bevatten in iedere denkbare vorm. Dit kan eenvoudige ASCII tekst zijn, maar ook afbeeldingen, tekstverwerkerbestanden, multimediabestanden, uitvoerbare programmabestanden of HTML.

Omdat bijlagen door e-mailclients vaak apart worden gepresenteerd, kunnen gebruikers de indruk hebben dat e-mail uit drie delen bestaat: header, body en attachment(s). Ook presenteren verschillende e-mailclients e-mails verschillend, waardoor gebruikers de indruk kunnen hebben dat e-mails qua formaat van elkaar kunnen verschillen. Dit is echter niet het geval: het e-mailbericht op het scherm is geenszins representatief voor de manier waarop het wordt verzonden. Het bestand dat verstuurd wordt, is in hoge mate gestandaardiseerd en bestaat uit platte tekst.

Het Technology Watch Report over E-mail Preservation van december 2011 van de Digital Preservation Coalition (http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf) meldt dat de Internet Engineering Taskforce (IETF, ietf.org) voor de uitwisseling van e-mailberichten het Simple Mail Transfer Protocol (SMTP, RFC 5321) heeft ontwikkeld. Deze standaard vereist dat berichten ASCII-gebaseerd worden verzonden. In RFC 5322 en de opvolgers daarvan heeft de IETF het formaat voor de bitstream beschreven: het Internet Message Format (IMF). Om ook niet-ASCII-gebaseerde berichten te kunnen versturen, wordt gebruik gemaakt van Multipurpose Internet Extensions (MIME). MIME voorziet in het coderen van niet-ASCII-tekens naar ASCII voor het verzenden van e-mailberichten, en het decoderen hiervan nadat het bericht is aangekomen. Vrijwel alle e-mailclients doen dit coderen en decoderen automatisch.

SMTP regelt hoe e-mailclients berichten naar e-mailservers sturen en hoe e-mailservers onderling e-mail uitwisselen. In het Internet Message Access Protocol (IMAP, momenteel IMAP4) en het Post Office Protocol (POP, momenteel POP3) wordt beschreven hoe mailclients e-mails van e-mailservers kunnen ophalen. In het algemeen verblijven bij IMAP de berichten op de server, terwijl bij POP de berichten door de client worden

bewaard. IMAP is ontwikkeld als opvolger van POP, maar beide protocollen worden nog aangeboden en gebruikt, omdat ze elk hun voor- en nadelen hebben.

Sommige e-mailclients hebben specifieke uitbreidingen op de e-mailstandaarden, maar allemaal ondersteunen ze ook de IETF-standaarden. De twee meest gangbare opslag- en uitwisselformaten zijn MBOX and EML. MBOX, soms het Berkeley-formaat genoemd, bestaat uit een set van vier licht verschillende opslagformaten en is oorspronkelijk ontwikkeld voor Unix-systemen. In het algemeen bestaat MBOX uit één bestand met de extensie .mbox of .mbx, en zit daarin de volledige inhoud van een e-mailfolder. Eventuele MIME-inhoud is in het bestand opgeslagen, waardoor MBOX-bestanden enorm groot kunnen worden. De kleinste corruptie kan er bij sommige e-mailclients voor zorgen dat een bericht (of zelfs de hele folder) onleesbaar wordt. Omdat MBOX-bestanden ook de bijlagen in hun MIME-format bevatten, is het waarschijnlijk dat ze op termijn gemigreerd moet worden t.b.v. duurzame toegankelijkheid van die bijlagen.

Bij EML-bestanden is doorgaans sprake van het opslaan van individuele e-mails als individuele bestanden. Bijlagen kunnen als MIME-inhoud worden opgeslagen in die bestanden, of als apart bestand waarnaar vanuit het EML-bestand gelinkt wordt.

Proprietaire e-mailclients kunnen berichten soms niet exporteren naar MBOX of EML. Het kan dan nodig zijn dat vanuit die clients wordt geëxporteerd naar andere, open of proprietaire, formaten. Bekende voorbeelden zijn lokale Outlook-gegevensbestanden (.pst), op Exchange servers opgeslagen e-maildatabases (.edb) of individueel opgeslagen Outlook items (.msg, "Outlook saved mail") van Microsoft, en Notes Storage Facility-bestanden (.nsf) van IBM Notes (voorheen Lotus Notes). Bij deze of andere, mogelijk uitgefaseerde e-mailopslagformaten, kan het nodig zijn de e-mails te migreren.

Meer informatie over e-mailpreservation:
- http://www.dpconline.org/component/docman/doc_download/739-dpctw11-01pdf (DPC Technology Watch Report over E-mail Preservation van december 2011)
- https://nl.wikipedia.org/wiki/E-mail (over e-mail)
- https://nl.wikipedia.org/wiki/Multipurpose_Internet_Mail_Extensions (over MIME)
- http://www.digitalpreservation.gov/formats/fdd/fdd000377.shtml (over .pst)

5.2 Risico-inventarisatie

5.2.1 Opslag

De protocollen en formaten voor het verzenden, transporteren en ontvangen van e-mail – het e-mailsysteem – zijn vrijwel allemaal gestandaardiseerd, open en goed gedocumenteerd. De grootste uitdagingen voor de duurzame toegankelijkheid van e-mail liggen aan de randen van het systeem: bij de opslag, de proprietaire opslagformaten, en de mate waarin die kunnen worden geëxporteerd naar gestandaardiseerde open formaten die duurzaam toegankelijk gemaakt en gehouden kunnen worden.

Bij het gebruik van MBOX-bestanden moet worden afgewogen in hoeverre het opslaan van alle e-mailinformatie (header, body en bijlagen) van alle e-mail in één bestand wenselijk is, vooral omdat een kleine beschadiging aan of corruptie in dat bestand een, meerdere of zelfs alle e-mails onleesbaar kan maken. EML-bestanden lijken dan een minder risicovol middel, zeker als bijlagen als afzonderlijke bestanden naast de e-mails worden bewaard.

5.3 Evaluatie

Dankzij (de facto) open en goed gedocumenteerde protocollen en standaarden m.b.t. e-mail, zoals IMF, MIME, SMTP, IMAP en POP, is e-mail in principe een goed ondersteund en goed te preserveren soort informatie.

Opgelet moet wel worden dat er (gedocumenteerde en/of open, maar tevens) proprietary e-mailformaten zoals .pst van Microsoft Outlook en bijv. .nsf van IBM zijn, die voor preservering t.z.t. geconverteerd moeten worden naar een open formaat. Er is nog geen alternatieve (de facto) standaard voor emailpreservering.

5.4 Alternatieven

Voor de opslag van email zijn diverse alternatieven, zoals opslag als XML, HTML, PDF of PDF/A. Op het moment van schrijven is er echter geen (de facto) standaard voor mailopslag, maar in de literatuur wordt overwegend aanbevolen EML als opslagformaat te gebruiken.

5.5 Ondersteuning in het e-Depot

5.5.1 Herkenning van Outlook-, Gmail- en Notes-mailboxen

Het e-Depot ondersteunt de ingest van Microsoft Outlook (.pst), Google Gmail (.mbox) en IBM Notes (.nsf). E-mails kunnen automatisch worden omgezet naar en opgeslagen als EML, waarbij iedere e-mail een eigen folder krijgt, waarin naast het e-mailbericht ook de evt. bijlagen zijn opgeslagen.

De metadata van de folder bevat metadata van de e-mail(header), zoals Van, Aan, Datum, Onderwerp en Bijlage(n). Op dit niveau bewaart het e-Depot ook de relatiegegevens van e-mail, zoals de volgende/vorige e-mail in een bepaalde e-maildiscussie.

Let op: als e-mail zou worden geconverteerd, dan wordt de e-mail eerst geïngest in het originele formaat (.pst, .mbox of bijv. .nsf) en daarna gemigreerd. Er zijn dan dus meteen twee (preservation)manifestaties: het originele opslagformaat en de EML-versie.

5.5.2 Formaten

In de Registry staan twee formaten met "message" in de naam, gaan twee ingangen over .pst-mailboxen en drie over .nsf:

PUID	Name	Version
fmt/278	Internet Message Format	
x-fmt/430	Microsoft Outlook Email Message	97-2003
x-fmt/248	Microsoft Outlook Personal Folders (ANSI)	1997-2002
x-fmt/249	Microsoft Outlook Personal Folders (Unicode)	2003-2007
x-fmt/336	Lotus Notes Database	2
x-fmt/337	Lotus Notes Database	3
x-fmt/338	Lotus Notes Database	4

5.5.3 Migration Pathways

De Registry kent een migratiepad voor het omzetten van e-mail:

PUID	Name	Version
TSS-pth/4	Apache POI MSG To EML	

5.5.4 Software en Tools

Voor het omzetten van e-mail noemt de Registry 1 tool:

PUID	Name	Version
sfw/10030	Apache POI MSG	3.9

5.6 Voorgestelde strategie

Het e-mailsysteem is een goed ondersteund en goed gedocumenteerd open systeem, waarbij vooral aan de opslagkant aandacht nodig is voor duurzame toegang.

Vanwege het presenteren (renderen) nu en de duurzame toegankelijkheid (preservation) later, ligt migratie naar EML, en het daarnaast apart opslaan van de bijlage(n), momenteel het meest voor de hand. Het originele bestand dient ook te worden bewaard. Deze strategie wordt ondersteund door het e-Depot.

Periodiek moet nagegaan worden in hoeverre nieuwe e-mailopslagformaten en –strategieën opkomen, zoals migratie naar XML of PDF(/A) resp. emulatie van emailsystemen. Dergelijke ontwikkelingen kunnen triggers zijn voor het aanpassen van deze strategie.

NB: op het moment van schrijven heeft het NA positief gereageerd op het toevoegen van EML aan de lijst open standaarden van het Forum Standaardisatie.

6 Portable Document Format (PDF)

Met behulp van Portable Document Format (PDF) kan een kleine of omvangrijke set tekstpagina's in één opgemaakt document over een netwerk uitgewisseld worden. Dit tekstformaat is afgeleid van PostScript en werd ontwikkeld door Adobe Systems Inc. Er wordt gebruik gemaakt van een markeringstaal om onder andere hotlinks, indexen en annotaties aan te brengen. Deze document-structurerende regels kunnen toegevoegd worden aan een bestaand PostScript-bestand met behulp van bijvoorbeeld het Distiller-programma dat onderdeel is van Adobe Acrobat. Voor het bekijken van een PDF-document is een PDF-viewer noodzakelijk.

De PDF bestanden waar we als NA op dit moment het meeste mee te maken krijgen kunnen onderverdeeld worden in drie hoofdgroepen:

- **Tekst-PDF** bestanden: De tekst in dit bestand wordt geëxtraheerd van een word document (dit kan ook een spreadsheet of een webpagina zijn) en in een PDF container geplaatst met details over de lettertypen, font grootte, locatie van de tekst etc. Het PDF bestand bevat de volledige tekst van het brondocument en is doorzoekbaar. Het PDF bestand bestaat uit één laag met tekst.
- **Image-PDF** bestanden: Een papieren document wordt gescand en daarna omgezet naar een PDF bestand. Het document bestaat uit een afbeelding van het originele papieren document wat tijdens het scannen is ontstaan en waarvan de tekst niet doorzoekbaar is. Het PDF bestand bestaat uit één laag met een afbeelding.
- **Image/OCR-PDF** bestanden: OCR software kan een grafische representatie (scan) van een letter of nummer lezen en omzetten naar tekst. Een scan van een papieren document (image) kan met OCR software omgezet worden naar een PDF bestand. Het bestand wat ontstaat bevat nog steeds de afbeelding (de scan) van de tekst maar heeft er een aparte laag met tekst informatie bij gekregen. Deze tekstlaag is doorzoekbaar. Het PDF bestand bestaat uit twee lagen.

Wanneer je deze verschillende PDF "groepen" in bijvoorbeeld een Windows Verkenner bekijkt zijn ze niet van elkaar te onderscheiden. Het is daarom belangrijk dat tijdens het creatieproces van het bestand/document goed vastgelegd wordt hoe en op welke manier het bestand gemaakt is. Identificatie tools die op dit moment in gebruikt zijn, zijn ook nog niet toereikend genoeg om exact te identificeren wat er in het PDF bestand aanwezig is.
NB: Het Europese VeraPDF project ontwikkelt op moment van schrijven een PDF validator tool die hierin moet voorzien. Zie http://verapdf.org/.

PDF veries 1.0 tot en met 1.6 zijn gesloten (proprietaire) formaten. PDF 1.7 is een open standaard en is door ISO en NEN genormeerd als ISO 32000-1:2008. Aan de nieuwe versie PDF 2.0 wordt op moment van schrijven van dit document nog gewerkt.

Er zijn diverse subsets van PDF:

- PDF/E: PDF/E staat voor PDF for Engineering en is een bestandsformaat voor het uitwisselen van engineering documentatie. De PDF/E standaard beschrijft hoe je op een betrouwbare manier engineering documenten kunt creëren, uitwisselen en beoordelen zelfs als het gaat om grote technische tekeningen. PDF/E heeft (net als PDF/A) richtlijnen of restricties waaraan de bestanden moeten voldoen.
- PDF/X: PDF/X staat voor PDF for Exchange en is gebaseerd op PDF v1.3, 1.4 en 1.6. Het doel van PDF/X is om uitwisselen van grafisch materiaal te vergemakkelijken. Het kent daarom een aantal afdruk-gerelateerde requirements die niet van toepassing zijn op standaard PDF bestanden.
- PDF/VT: PDF for Exchange of Variable Data and Transactional (VT) Printing. Gebaseerd op PDF 1.4 of 1.6.
- PDF/UA: PDF for Universal Accessibility (gebaseerd op PDF 1.7) is eigenlijk een set van richtlijnen voor het creëren van de zogenaamd "toegankelijke" PDF bestanden. Het formaat is ontworpen om elk soort elektronisch document (graphics, tekst, multimedia, audio) in PDF toegankelijk te maken. De doelgroep waar dit bestandsformaat zich op richt zijn voornamelijk mensen met een handicap (text-to-speech screen readers voor slechtzienden bijvoorbeeld).
- PDF/A (zie volgende hoofdstuk) is een subset van het PDF formaat, die is ontwikkeld voor de langetermijnarchivering van documenten.

  In tegenstelling tot PDF/A heeft PDF 1.7:[12]
  - wel klikbare links
  - mag encrypted zijn
  - heeft wachtwoorden
  - wel transparancy
  - wel geluid
  - wel video
  - wel 3D

PDF 1.7 is opgenomen op de ‘pas-toe-of-leg-uit’ lijst met open standaarden van het Forum Standaardisatie.[13]

In de rest van dit document wordt, tenzij anders aangegeven, ingezoomd op het speciaal voor archiveringsdoeleinden ontwikkelde PDF/A formaat.

[12] Bron: http://www.den.nl/standaard/36/Portable-Document-Format (geraadpleegd 22-10-2015)
[13] Bron: https://lijsten.forumstandaardisatie.nl/open-standaard/pdf-17 (geraadpleegd 19-11-2015)

7 Portable Document Format Archivable (PDF/A)

7.1 Algemene informatie

Bestandsformaat ontworpen voor langetermijnarchivering (long term preservation) van elektronische tekstdocumenten, inclusief raster, vector en andere data. [14]

Beherende organisatie: ISO
PDF/A-1:
http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=38920
PDF/A-2:
http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50655
PDF/A-3:
http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229

Overige informatiebronnen:
ISO 19005-1:2005-12 en Documentbeheer - Elektronische bestandsformaat document voor langdurige bescherming - Gebruik van PDF 1.4 (PDF/A-1)
Document management -- Electronic document file format for long-term preservation -- Part 2: Use of ISO 32000-1 (PDF/A-2)
NEN-ISO 19005-3:2012 en Documentbeheer - Elektronische bestandsformaat document voor langdurige bescherming - Deel 3: Gebruik van de ISO 32000-1 met ondersteuning voor ingesloten bestanden (PDF/A-3)
Beschrijving PDF/A in Library of Congres register van duurzame formaten
Wikipedia PDF/A lemmaArtikel over het gebruik van PDF/A op de project CEST website
The NDSA rapport: Benefits and Risks of the PDF/A-3 file Format For Archival Institutions
PDF/A Competence Center
PDF Guidelines, Recommendations for the creation of PDF files for long-term preservation and access (Rapport van de Koninklijke Bibliotheek)
http://www.opf-labs.org/format-corpus/pdfCabinetOfHorrors/ (PDF Cabinet of Horrors on OPF Format Corpus)

7.1.1 PDF/A-1

In tegenstelling tot het reguliere PDF formaat kent PDF/A een aantal restricties waardoor het eenvoudiger te bewaren is. PDF/A-1 kan onafhankelijk van hard- of softwareplatformen betrouwbaar en consistent worden weergegeven. Dit komt omdat het formaat 'self-contained' is, of te wel: het bevat alle bronnen, met name fonts, om betrouwbare weergave mogelijk te maken. Er zijn twee niveaus waaraan kan worden voldaan:

a) PDF/A-1a: Voldoet aan volledige eisen van de standaard. De tekst is naast correcte weergave ook doorzoekbaar (tekst is gecodeerd als Unicode). De logische structuur (koppen, paragrafen etc) van tekst zijn bewaard. Ook wel 'tagged PDF'

[14] Bron: http://www.den.nl/standaard/168/Portable-Document-Format-Archivable (geraadpleegd 22-10-2015)

b) PDF/A-1b: Voldoet aan minimale eisen van de standaard. Tekst (en andere content) worden correct weergegeven maar de tekststructuur ontbreekt. Dit formaat wordt gebruikt voor gescande documenten (al dan niet voorzien van een OCT tekstlaag), elektronisch geboren tekst zonder structuurelementen of tekst die is opgemaakt in oudere software. Wanneer PDF/A via een printer driver wordt gegenereerd is deze altijd PDF/A-1b.

De eerste versie van PDF/A is uitgebracht in 2005 als PDF/A-1 gebaseerd op PDF 1.4. In juli 2011 is de tweede versie van de standaard gepubliceerd: PDF/A-2. In 2012 volgde versie 3: PDF/A-3. De twee laatsten zijn gebaseerd op PDF 1.7.

7.1.2 PDF/A-2

PDF/A-2 is een aanvulling op PDF/A-1 waarbij PDF/A-1 documenten ook ondersteund worden door PDF/A-2, maar PDF/A-2 documenten niet altijd ondersteund worden door PDF/A-1.

PDF/A-2 is gebaseerd op PDF1.7 en kan gezien worden als een subset van deze ISO-standaard en is compatibel met ander ISO standaarden zoals PDF/X-4 en PDF/E-1. Voor alle duidelijkheid: PDF/A-2 is geen "upgrade" van PDF/A-1. Als documenten geen gebruik maken van de nieuwe functionaliteit die in PDF 1.7 zitten (en natuurlijk ondersteund worden door PDF/A-2) dan kan er nog steeds voldaan worden met PDF/A-1. De PDF/A-2 standaard vervangt dus niet PDF/A-1. Het doel van PDF/A-2 is een mogelijkheid bieden om op een betrouwbare manier de "nieuwe" functies te archiveren zonder dat daarbij concessies gedaan hoeven te worden aan de inhoud. Bovenop PDF/A-1 biedt PDF/A-2 onder andere de volgende mogelijkheden:

- JPEG2000 beeldcompressie.
- Ondersteuning van transparantie en lagen.
- Ondersteuning van OpenType lettertypes.
- Gebruik van digitale handtekeningen volgens PAdES (PDF Advanced Electronic Signatures).
- Mogelijkheid om meerdere documenten samen te voegen en deze als individuele documenten op te slaan binnen één bestand.

7.1.3 PDF/A-3

In 2012 is versie 3 van PDF/A uitgebracht. Meest opvallende en omstreden uitbreiding van het formaat is de mogelijkheid om andersoortige bestanden in te sluiten in de PDF schil. Zo kan bijvoorbeeld een rekenblad, een tekstverwerkingsbestand of een CSV bestand worden ingesloten. Er wordt hierbij wel gesproken van "hybride archivering" of een gebundeld formaat. Een belangrijk nadeel hiervan is dat zo'n PDF/A niet langer 'self-contained' is. Er zijn immers andere applicaties nodig om de ingesloten bestanden te kunnen uitlezen. De status van PDF/A-3 voor lange termijn archivering is daarmee omstreden. Er zijn wellicht use cases denkbaar waarin gebundelde documenten nuttig kunnen zijn - voorbeelden daarvan zijn te vinden in de omschrijving van PDF/A-3 door de Library of Congres. In het 2014 verschenen NSDA (National Digital Stewardship Alliance) rapport *The Benefits and Risks of the PDF/A-3 file Format For Archival Institutions* wordt - met zoveel woorden – het gebruik van PDF/A-3 voor archivering ontraden. Mocht er grote behoefte bestaan aan de fysieke bundeling van bestanden

dan liggen formaten als BagIt File Packaging Format en (een beperkte vorm van) ZIP meer voor de hand.

7.2 Risico-inventarisatie

7.2.1 Embedded files in PDFA/3 [15]

Het uitgangspunt bij PDF/A is dat het primaire document van een PDF/A bestand voldoende beschermd is tegen preserveringsrisico's op de (zeer) lange termijn. Echter, in het geval van PDF/A-3 is er geen vereiste dat eventuele ingesloten (embedded) bestanden ook archiefwaardig zijn. Een PDF/A-3 conforme reader zorgt allereerst voor het presenteren van het primaire document, maar maakt ook extractie mogelijk van ingesloten bestanden voor gebruik met andere tools.
Er zijn scenario's denkbaar waarin een dergelijke 'archief bundeling " van PDF/A-3 misschien zinvol is. Voorwaarde voor geheugeninstellingen voor een dergelijk gebruik van embedded bestanden in PDF/ A-documenten is, dat er specifieke afspraken worden gemaakt tussen archiefvormers en archiefinstellingen, waarbij de toegestane ingesloten bestandsformaten duidelijk worden benoemd, en waarbij een workflow wordt gedefinieerd die garandeert dat de relatie tussen het PDF-document en eventuele embedded files volkomen duidelijk is voor de archiefinstelling.
Van grotere zorg is de mogelijkheid dat PDF/A-3 in sommige gevallen wordt gebruikt als een algemeen "bundelingsformaat", waarbij het zichtbare primaire document van minder langdurige belang is dan de secundaire, ingesloten bestanden.

7.2.2 Conversie van PDF naar PDF/A [16]

In theorie zijn er goede redenen te bedenken waarom conversie van PDF naar PDF/A nuttig kan zijn. Zo kunnen externe afhankelijkheden een PDF kwetsbaar maken voor incompatibele of minder authentieke weergave - bij een missend font bijvoorbeeld - van de content. De huidige praktijk is echter anders. De conversie van PDF naar PDF/A is foutgevoelig en er zijn - ernstig genoeg - geen goede validators voorhandig om de mogelijke fouten te achterhalen.[17]

7.2.3 Verminkte (not valid) PDF [18]

Sommige applicaties produceren PDF’s die niet voldoen aan de PDF formaat specificaties (PDF 1.7 /ISO 32000-1 of de eerdere pre-ISO specificaties). PDF zou in dat geval niet correct (of in het geheel niet) kunnen renderen. Toekomstige migraties naar alternatieve formaten kunnen resulteren in dataverlies of kunnen helemaal mislukken. Validatie is problematisch voor PDF, vooral vanwege de complexiteit van het formaat en het gebrek aan betrouwbare tools.

7.2.4 Encryptie [19]

PDF staat het gebruik van encryptie toe om toegang of hergebruik van de inhoud van een document te beperken. Dit varieert van documenten die alleen met behulp van een wachtwoord geopend kunnen worden tot het uitzetten van bepaalde functionaliteit (bijvoorbeeld afdrukken of content

[15] Bron: http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_PDF_A3_report_final022014.pdf
[16] Bron: http://www.den.nl/standaard/168/Portable-Document-Format-Archivable (geraadpleegd 22-10-2015)
[17] Lees meer over de moeizame PDF naar PDF/A conversie in de Open Planets blog When (not) to migrate a PDF to PDF/A.
[18] Bron: http://wiki.opf-labs.org/display/TR/Not+valid+PDF (geraadpleegd 22-10-2015)
[19] Bron: http://wiki.opf-labs.org/display/TR/Encryption (geraadpleegd 22-10-2015)

kopiëren). Content kan onleesbaar worden als het wachtwoord onbekend is (hoewel het "kraken" van wachtwoorden vaak technisch wel mogelijk is, zitten daar juridisch gezien soms haken en ogen aan). Restricties op bijvoorbeeld afdrukken en kopiëren kunnen toekomstige preserveringsacties bemoeilijken.

7.2.5 Ontbrekende, beschadigde of incomplete fonts [20]

Een veel voorkomende categorie van problemen is dat PDFs gebruik kunnen maken van fonts die niet aan het bestand zijn toegevoegd (embedded), of beschadigd of incompleet zijn. Als gevolg hiervan kunnen PDF's niet correct renderen. Er zijn nog veel meer issues met fonts, die in grofweg in drie categorieën kunnen worden gevat:

- Ongeldige of incomplete *dictionary* fouten. Deze categorie omvat een reeks problemen, inclusief fonts die niet zijn toegevoegd (embedded).
- Beschadigde embedded font fouten.
- Glief (glyph) fouten.

7.2.6 JavaScript [21]

PDF documenten kunnen JavaScript bevatten wat tot beveiligingsproblemen kan leiden.

7.2.7 Verwijzingen naar externe documenten [22]

Een PDF kan verwijzingen bevatten naar externe documenten. Ieder extern document waarnaar in de PDF wordt verwezen kan echter veranderen of verdwijnen, waardoor het renderen van het PDF bestand kan worden beïnvloed. Zelfs in PDF/A bestanden kunnen externe verwijzingen worden opgenomen, volgens de volgende mechanismen:

- URI actions: dit zijn verwijzingen naar het Internet (bijvoorbeeld een klikbare hyperlink).
- GoToR acties: deze verwijzen naar een extern PDF bestand (bijvoorbeeld een klikbare link naar een lokaal opgeslagen PDF bestand)

Vanuit preservation standpunt vormen GoToR acties een risico (bijvoorbeeld in geval van een collective PDF's die naar elkaar verwijzen), ook al wordt het renderen van de bestanden die de verwijzing bevatten niet beïnvloed.[23]

7.2.8 Bestandsbijlagen[24]

PDFs kunnen bestandsbijlagen bevatten. Deze bijlagen kunnen op twee manieren worden toegevoegd aan een PDF:

1. Bijlagen op paginaniveau die gebruik maken van een File Attachment Annotation (sectie 12.5.6.15 of ISO32000)
2. Bijlagen op documentniveau die worden gedefinieerd door de EmbeddedFiles entry in de name dictionary van het document (sectie 7.7.4 of ISO32000)

Beide zijn slechts referenties naar de eigenlijke data van de bestandsbijlagen, die in beide gevallen als een Embedded File Stream in het

[20] Bron: http://wiki.opf-labs.org/display/TR/Fonts+missing%2C+damaged+or+incomplete (geraadpleegd 22-10-2015)
[21] Bron: http://wiki.opf-labs.org/display/TR/JavaScript (geraadpleegd 22-10-2015)
[22] Bron: http://wiki.opf-labs.org/display/TR/References+to+external+files (geraadpleegd 22-10-2015)
[23] Meer informatie is te vinden in deze whitepaper: http://www.ebriefpro.com/pdfs/pdfa.pdf (geraadpleegd 07-12-2015)
[24] Bron: http://wiki.opf-labs.org/display/TR/File+attachments (geraadpleegd 22-10-2015)

document zijn opgeslagen. NB: een Embedded File Stream kan ook worden gebruikt om multimedia content te representeren, dus uitsluitend daarmee kan een bestandsbijlage niet worden geïdentificeerd. Bijlagen kunnen ieder willekeurig formaat hebben, waardoor er risico's zijn op het gebied van duurzame toegang. En bijlagen met kwaadaardige software bevatten vormen een beveiligingsrisico.

7.2.9 Multimedia content [25]

PDFs kunnen multimedia content bevatten: filmpjes, geluidsfragmenten en 3D content. Rendering van multimedia content kan afhankelijk zijn van externe applicaties die in de toekomst misschien niet beschikbaar zijn; daarnaast kan het formaat waarin de multimedia content is opgeslagen verouderd raken.

7.3 Evaluatie

PDF/A is een wijd verbreide open standaard, een NEN/ISO norm (ISO:19005). Gezien de bovengeschetste risico's m.b.t. embedded files is het gebruik van versie PDF/A3 niet aan te bevelen.
Ook zijn er nogal wat andere risico's die aandacht vereisen. Bovenaan staan hierbij PDF bestanden met encryptie en het risico van de niet-ingesloten fonts.

7.4 Alternatieven

PDF v1.7 (ISO32000-1). Oudere PDF versies 1.0 tot en met 1.6 zijn gesloten, proprietaire formaten).
Er zijn verschillende gratis en vrije alternatieven beschikbaar voor PDF zoals LibreOffice, OpenOffice.org en AbiWord.[26]

7.5 Ondersteuning in het e-Depot

7.5.1 Formaten

In de registry voldoen negenentwintig formaten aan de zoekterm "pdf":

PUID	Name	Version
fmt/14	Acrobat PDF 1.0 - Portable Document Format	1.0
fmt/15	Acrobat PDF 1.1 - Portable Document Format	1.1
fmt/16	Acrobat PDF 1.2 - Portable Document Format	1.2
fmt/17	Acrobat PDF 1.3 - Portable Document Format	1.3
fmt/18	Acrobat PDF 1.4 - Portable Document Format	1.4
fmt/19	Acrobat PDF 1.5 - Portable Document Format	1.5
fmt/20	Acrobat PDF 1.6 - Portable Document Format	1.6
fmt/276	Acrobat PDF 1.7 - Portable Document Format	1.7
fmt/354	Acrobat PDF/A - Portable Document Format	1b
fmt/476	Acrobat PDF/A - Portable Document Format	2a
fmt/477	Acrobat PDF/A - Portable Document Format	2b
fmt/478	Acrobat PDF/A - Portable Document Format	2u
fmt/479	Acrobat PDF/A - Portable Document Format	3a
fmt/480	Acrobat PDF/A - Portable Document Format	3b
fmt/481	Acrobat PDF/A - Portable Document Format	3u
fmt/95	Acrobat PDF/A - Portable Document Format	1a

[25] Bron: http://wiki.opf-labs.org/display/TR/Multimedia+content (geraadpleegd 22-10-2015)
[26] Bron: https://nl.wikipedia.org/wiki/Microsoft_Word (geraadpleegd 16-11-2015)

PUID	Name	Version
fmt/493	Acrobat PDF/E - Portable Document Format for Engineering PDF/E-1	
fmt/144	Acrobat PDF/X - Portable Document Format - Exchange 1:1999	
fmt/145	Acrobat PDF/X - Portable Document Format - Exchange 1:2001	
fmt/157	Acrobat PDF/X - Portable Document Format - Exchange 1a:2001	
fmt/146	Acrobat PDF/X - Portable Document Format - Exchange 1a:2003	
fmt/147	Acrobat PDF/X - Portable Document Format - Exchange 2:2003	
fmt/158	Acrobat PDF/X - Portable Document Format - Exchange 3:2002	
fmt/148	Acrobat PDF/X - Portable Document Format - Exchange 3:2003	
fmt/488	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4	
fmt/489	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4p	
fmt/490	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5g	
fmt/492	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5n	
fmt/491	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5pg	

NB: PDF/VT en PDF/UA komen niet voor in de Registry. Het is echter niet ondenkbaar dat deze formaten op een gegeven moment door Vormers worden aangeboden.

7.5.2 Migration pathways

Het e-Depot kent maar liefst 128 migration pathways om alle mogelijke formaten om te zetten naar PDF v1.3, v1.4 of v1.5 (NB: dus geen hogere versies). Daarnaast zijn er 91 migration pathways voor omzetting naar PDF/A. Hieronder zijn ook migration pathways om PDF bestanden (v1.0 – 1.7) te migreren naar het PDF/A formaat.

PUID	Name	Version
pth/117	BFO PDF 1.0 to PDF/A	
pth/118	BFO PDF 1.1 to PDF/A	
pth/119	BFO PDF 1.2 to PDF/A	
pth/120	BFO PDF 1.3 to PDF/A	
pth/121	BFO PDF 1.4 to PDF/A	
pth/122	BFO PDF 1.5 to PDF/A	
pth/123	BFO PDF 1.6 to PDF/A	
pth/297	BFO PDF 1.7 to PDF/A	

7.5.3 Software en Tools

De beschikbare software voor het omgaan met PDF bestaat uit:

PUID	Name	Version

PUID	Name	Version
sfw/10006	Big Faceless PDF Library	2.15.2
sfw/2	LuraDocument PDF Compressor	
sfw/10027	PdfTron PDF/A Manager	6.2

Er zijn 12 PDF tools beschikbaar:

PUID	Name	Version
tool/1003	BFO Image To PDF/A	
tool/101	BFO PDF To PDF/A	
tool/101250	BFO Validate PDF	
tool/101284	BFO Validate PDF/A	
tool/102587	Image Magick Convert to PDF	
tool/106059	Jhove-Pdf	
tool/106087	Jhove-PdfA	
tool/106999	Open Office PDF	
tool/107	Open Office PDF/A	
tool/107170	PDFTRON Validate PDFA	
tool/109	Stellent PDF Export	
tool/109373	Uniconvertor convert to PDF	

De standaard e-Depotcomponenten identifyFileComponents en identifyWebpages (her)kennen PDF/A en PDF 1.0 t/m 1.7. Ook hieruit blijkt de brede ondersteuning van PDF.

7.6 Voorgestelde strategie

PDF/A-1 en PDF/A-2 zijn opgenomen op de 'pas-toe-of-leg-uit' lijst met open standaarden van het Forum Standaardisatie (zie ook paragraaf over PDF).[27] Er bestaan (nog) geen algemeen geaccepteerde tools voor PDF validatie (op termijn: VeraPDF?), maar gebruik van de tool Apache Preflight en controle van de output daarvan op fouten zal op zijn minst PDF's detecteren die ernstige gebreken vertonen. In sommige gevallen is het misschien mogelijk om via de Vormer een intacte versie van verminkte documenten te verkrijgen.[28]

Verder:

- Vanwege de foutgevoeligheid liever geen conversies van PDF naar PDF/A, zeker niet door de Vormer zelf. Mocht de conversie als absoluut noodzakelijk worden gezien dan wordt in ieder geval aangeraden de originele PDF te bewaren.
- Laat geen PDF bestanden toe die gebruik maken van encryptie.
- Voor duurzame toegang van bijlagen en embedded multimedia content kan het een optie zijn om de content te extraheren en als apart bestand bij de PDF (supplementary file object) te ingesten.[29]

[27] Bron: https://lijsten.forumstandaardisatie.nl/ (geraadpleegd 23-11-2015)
[28] Bron: http://wiki.opf-labs.org/display/TR/Not+valid+PDF (geraadpleegd 22-10-2015)
[29] Bron: http://wiki.opf-labs.org/display/TR/Multimedia+content (geraadpleegd 22-10-2015)

8 MS Office algemeen

Word, Excel, Powerpoint en Access zijn MS Office-programma's die in dit document worden beschreven. Sinds 2007 hebben Word, Excel en Powerpoint een gedeeld bestandsformaat. Dit Office Open XML-formaat (OOXML) "*is a zipped, XML-based file format developed by Microsoft for representing spreadsheets, charts, presentations and word processing documents*." (https://en.wikipedia.org/wiki/Office_Open_XML). Word, Excel en Powerpoint worden in de volgende hoofdstukken apart beschreven, maar hebben, historisch gezien en vanwege hun gedeelde bestandsformaat, ook algemene kenmerken. Die algemene kenmerken beschrijven we in dit hoofdstuk. Access wordt ook apart beschreven, maar maakt geen gebruik van OOXML.

Microsoft ontwikkelde OOXML in 2007. Het is een op XML gebaseerd formaat en daarmee een breuk met de oudere versies van de Office suite, die waren gebaseerd op proprietaire, binaire bestandformaten. Sinds Office 2007 is OOXML het standaard bestandsformaat voor MS Office. De ISO standaard voor OOXML is ISO/IEC DIS 29500.[30] Een OOXML-bestand bestaat uit een gecomprimeerd ziparchief dat wordt aangeduid afhankelijk van het bestandstype (.docx, .xlsx, .pptx). OOXML is platformonafhankelijk.

OOXML gebruikt drie speciale (custom) op XML gebaseerde talen om drie document typen te beschrijven: WordProcessingML, SpreadsheetML and PresentationML worden gebruikt voor respectievelijk Word, Excel en Powerpoint. OOXML werd ontwikkeld als antwoord op de groeiende vraag naar officesoftware met open bestandsformaten. Het is een goed gedocumenteerd open formaat waar oudere versies van Officedocumenten vrijwel zonder verlies van informatie naar toe gemigreerd kunnen worden. Er is bijv. enige controverse over het standaardisatieproces van deze open standaard, zie bijv. https://nl.wikipedia.org/wiki/Office_Open_XML. De specificaties en de ontwikkelrechten zijn wel vrijgegeven onder Microsofts Covenant not to sue.

Er zijn verschillende commerciële, gratis en vrije alternatieven beschikbaar voor de MS Office suite, zoals LibreOffice, OpenOffice.org en AbiWord.[31] Opgemerkt moet echter worden dat MS Office op het moment van schrijven de dominante marktleider is.

Voor het omzetten van oudere Officebestanden naar OOXML levert Microsoft Compatibility Packs. Deze ondersteunen "some versions", maar het is onduidelijk hoeveel en welke eerdere versies er dan precies ondersteund worden. Laten we er vanuit gaan dat de Compatibility Packs maximaal 3 versies terug gaan. De laatste Compatbility Pack for Office 2007 Systems ondersteunt MS Word 97-2003 bestanden (en niet lager). Microsoft maakt gebruik van "life cycle support" (https://support.microsoft.com/nl-nl/gp/lifeselect) en op deze website kun je vinden welke Microsoft producten nog actief worden ondersteund.

[30] Bron: Preservica 5.4 registry https://e-Depot-acpt.nationaalarchief.nl/Registry/registry.html#list:fmt/189&false (geraadpleegd 07-12-2015)

[31] Bron: https://nl.wikipedia.org/wiki/Microsoft_Word (geraadpleegd 16-11-2015)

Risico-inventarisaties, evaluaties en (preservation)strategieën voor de afzonderlijke Officeprogramma's worden beschreven in de volgende hoofdstukken.

9 MS Word

9.1 Algemene informatie

Microsoft Word, of meestal alleen Word, is een van de meest gebruikte tekstverwerkers ter wereld. Het is ontwikkeld door Microsoft, als onderdeel van de MS Office suite. Tot in het begin van de jaren 90 was WordPerfect de meest gebruikte tekstverwerker, maar toen voor de pc het DOS-besturingssysteem door Windows werd vervangen, werd Word de algemeen gebruikte tekstverwerker. Word is in veel talen beschikbaar.

Microsoft Word werkt op Windows en OS X platformen. Met behulp van Wine werken versies pre-2013 ook op Linux.[32] Ook op de mobiele besturingssystemen van Apple (iOS) en Google (Android) en natuurlijk Microsoft zelf (Windows) wordt MS Word ondersteund.

Overige informatiebronnen:
http://www.digitalpreservation.gov/formats/fdd/fdd000397.shtml (over .docx transitional) en
http://www.digitalpreservation.gov/formats/fdd/fdd000400.shtml (over .docx strict).
https://en.wikipedia.org/wiki/Microsoft_Word

De bestandsnaam extensie van Microsoft Word's document formaten is .doc or .docx. Hoewel de .doc extensie in veel verschillende versies van Word is gebruikt omvat deze in feite vier verschillende bestandsformaten:

1. Word for DOS
2. Word for Windows 1 en 2: Word 3 en 4 for Mac OS
3. Word 5 en Word 95 for Windows; Word 6 for Mac OS
4. Word 97 en later for Windows; Word 98 en later for Mac OS

De nieuwere .docx extensie is geassocieerd met de OOXML standaard en wordt gebruikt door Word 2007, 2010 en 2013 for Windows, Word 2008 en 2011 for Mac OS X, alsook door een groeiend aantal applicaties van andere firma's, inclusief OpenOffice.org Writer, een open source tekstverwerker.[33]

9.2 Risico-inventarisatie

9.2.1 Gesloten (proprietair) formaat

Met de release van Word 6.0 introduceerde Microsoft een nieuw eigen binair tekstverwerkingsformaat, gebaseerd op het generieke OLE2 Compound Document Format. Het formaat is proprietair en Microsoft maakt de details van de structuur niet openbaar. Het formaat bleef onveranderd met de release van Word 95, 97, 2000, 2002 en 2003.[34] Zoals gezegd is vanaf

[32] Bron: https://nl.wikipedia.org/wiki/Microsoft_Word (geraadpleegd 16-11-2015)
[33] Bron: https://en.wikipedia.org/wiki/Microsoft_Word (geraadpleegd 24-11-2015)
[34] Bron: Preservica 5.4 registry https://e-Depot-acpt.nationaalarchief.nl/Registry/registry.html#list:fmt/39&false en https://e-Depot-acpt.nationaalarchief.nl/Registry/registry.html#list:fmt/40&false (geraadpleegd 24-11-2015)

Microsoft Office 2007 het standaard output formaat van MS Word gebaseerd op het meer open OOXML file formaat.[35]

9.2.2 Interoperabiliteit tussen de verschillende Office Versies

Op de MS Office support website[36] zijn er referenties te vinden van de verschillende Word versies. Deze referenties zijn gebaseerd op de functionaliteit die per versie kan veranderen. Er wordt een onderverdeling gemaakt tussen MS Word 97-2003, MS Word 2007 en MS Word 2010. Over eerdere versies is weinig concreets te vinden.

In alle interoperabiliteit vergelijkingen tussen de verschillende Word versies worden de versies die gemaakt zijn tussen 97-2003 als één groep gezien. Ook DROID identificeert deze documenten als groep (Microsoft Word for Windows Document 97-2003). Toch kunnen er problemen optreden met het openen van .doc bestanden in oudere of nieuwere versies van MS Word. In documenten over bestandsformaten wordt beschreven dat elke MS Word versie zijn eigen bestandsformaat hanteert.

9.2.3 Object Linking and Embedding (OLE) Data Structures

Door het gebruik van de OLE Data Structuur is het mogelijk voor applicaties om documenten te creëren die 'gelinkte' of 'ingesloten' objecten bevatten. Je kunt hierbij een onderscheid maken tussen de 'container' applicatie en de 'creating' applicatie. Het bestandsformaat voor een ingesloten object is anders dan het bestandsformaat voor een gelinkt object. Het ingesloten object moet zowel de oorspronkelijke data als de data over de applicatie waarmee het gecreëerd is bevatten. Het gelinkte object daarentegen hoeft alleen een referentie naar deze data te bevatten. Beide formaten bevatten data die nodig is om het gelinkte of ingesloten object te kunnen weergeven binnen de container applicatie.

9.3 Evaluatie

Hoe goed gedocumenteerd, gestandaardiseerd en vrij van licenties/patenten ook, het Word-formaat is een formaat dat eigendom is van en/of alleen bewerkt kan worden door Microsoft, en daarmee geen volledig open formaat. De meest recente OOXML-versies vormen een beperkter preservationrisico.

9.4 Alternatieven

Zie voor overzichten van alternatieve tekstverwerking software:
- https://en.wikipedia.org/wiki/List_of_word_processors
- http://alternativeto.net/software/microsoft-word/

Daarnaast zijn het PDF en PDF/A formaat alternatieven voor het .doc en .docx formaat. Beide zijn open standaarden die zijn opgenomen in de 'pas-toe-of-leg-uit' lijst van het Forum Standaardisatie.

[35] Bron: Preservica 5.4 registry https://e-Depot-acpt.nationaalarchief.nl/Registry/registry.html#list:fmt/412&false (geraadpleegd 24-11-2015)

[36] Zie: https://support.office.com/en-US/article/Compatibility-changes-between-versions-CB713C85-3145-4E83-A8B6-1E3227A4C059 (geraadpleegd 26-11-2015)

9.5 Ondersteuning in het e-Depot

9.5.1 Formaten

Zoeken naar "Microsoft Word" in de registry levert 20 verschillende formaten op:

PUID	Name	Version
fmt/346	Microsoft Word for Macintosh Document	1.0
fmt/37	Microsoft Word for Windows Document	1.0
fmt/38	Microsoft Word for Windows Document	2.0
fmt/39	Microsoft Word Document	6.0/95
fmt/40	Microsoft Word Document	97-2003
fmt/412	Microsoft Word for Windows	2007 onwards
fmt/523	Macro enabled Microsoft Word Document OOXML	2007 onwards
fmt/597	Microsoft Word Template	2007 onwards
fmt/599	Microsoft Word Macro-Enabled Document Template	2007 onwards
x-fmt/1	Microsoft Word for Macintosh Document	3.0
x-fmt/129	Microsoft Word for Macintosh Document	X
x-fmt/2	Microsoft Word for Macintosh Document	6.0
x-fmt/204	Microsoft Word for Windows Macro	
x-fmt/273	Microsoft Word for MS-DOS Document	3.0
x-fmt/274	Microsoft Word for MS-DOS Document	4.0
x-fmt/275	Microsoft Word for MS-DOS Document	5.0
x-fmt/276	Microsoft Word for MS-DOS Document	5.5
x-fmt/45	Microsoft Word Template	
x-fmt/64	Microsoft Word for Macintosh Document	4.0
x-fmt/65	Microsoft Word for Macintosh Document	5.0

9.5.2 Migration pathways

Het e-Depot kent diverse migration pathways om verschillende versies van MS Word bestanden om te zetten naar een ander formaat:

PUID	Name	Version
pth/19	Microsoft Word for Windows 1.0 to HTML	
pth/35	Microsoft Word for Windows 1.0 to Portable Document Format 1.4	
pth/20	Microsoft Word for Windows 2.0 to HTML	
pth/36	Microsoft Word for Windows 2.0 to Portable Document Format 1.4	
pth/2	Microsoft Word for Windows 2.0 to XML 1.0	
pth/21	Microsoft Word for Windows 6.0/95 to HTML	
pth/37	Microsoft Word for Windows 6.0/95 to Portable Document Format 1.4	
pth/3	Microsoft Word for Windows 6.0/95 to XML 1.0	
pth/22	Microsoft Word for Windows 97-2003 to HTML	
pth/34	Microsoft Word for Windows 97-2003 to Portable Document Format 1.4	
pth/1	Microsoft Word for Windows 97-2003 to XML 1.0	

NB: er is geen workflow waarmee oude .doc bestanden kunnen worden gemigreerd naar het .docx formaat.

9.5.3 Software en Tools

Beschikbare software voor het omgaan met MS Word bestaat uit:

PUID	Name	Version
x-sfw/1	Word	97 (8.0) for Windows
x-sfw/10	Word	95 (7.0) for Windows
x-sfw/11	Word	6.0 for Windows
x-sfw/12	Word	2.x for Windows
x-sfw/14	Word	1.0 for Windows
x-sfw/15	Word	4.0 for Macintosh
x-sfw/16	Word	5.0 for Macintosh
x-sfw/3	Word	2002/XP (10.0) for Windows
x-sfw/43	Word	X
x-sfw/17	Word	98 for Macintosh
x-sfw/18	Word	2001 for Macintosh
x-sfw/19	Word	X for Macintosh
x-sfw/2	Word	2000 (9.0) for Windows
x-sfw/269	Word	2003
x-sfw/3	Word	2002/XP (10.0) for Windows
x-sfw/43	Word	X

Er zijn in Preservica geen tools beschikbaar specifiek voor Word (.doc en .docx).

9.6 Voorgestelde strategie

De voorgestelde strategie voor Wordbestanden is dat momenteel geen acties noodzakelijk zijn, omdat sommige Wordformaten weliswaar verouderd, maar nog altijd leesbaar zijn en voldoende door software ondersteund worden.

Periodiek moet nagegaan worden of deze strategie aangepast moet worden. Omdat Microsoft marktleider is op het gebied van tekstverwerkers, alternatieve tekstverwerkers (met een open formaat) soms minder ver uitontwikkeld zijn, en er informatieverlies kan optreden bij migratie naar andere formaten, zijn er dan meerdere mogelijkheden:

- migratie van Word-bestanden naar PDF/A bestanden
- migratie van Word-bestanden naar het OpenDocument-formaat voor tekstdocumenten (.odf)
  Let op: volledig behoud van informatie (data, functionaliteit en layout) is verre van verzekerd en de kwaliteit moet van geval tot geval gecontroleerd worden!
- als volledig behoud van informatie (data, functionaliteit en layout) noodzakelijk is en de ondersteuning van het (ver)oude(rde) Word-formaat onvolledig is of stopt: periodieke conversie van (ver)oude(rde) Word-bestanden naar de meest recente versie, momenteel OOXML

In alle gevallen heeft het de voorkeur het originele Word-bestand te bewaren, omdat er in de toekomst nieuwe mogelijkheden voor preservering kunnen ontstaan.

10 MS Office Excel

10.1 Algemene informatie

Een spreadsheet bestaat uit interactieve tabellen waarin elk data item (cell) een formule kan bevatten en daardoor dynamisch gelinkt is aan een ander data item waardoor de inhoud kan veranderen. De meeste programma's voor spreadsheet voegen verschillende werkbladen (worksheet) samen tot één bestand. Binnen het bestand (en zelfs naar andere bestanden) kunnen aparte werkbladen gelinkt zijn door een formule. Spreadsheets kunnen gezien worden als tabellen met extra functionaliteiten. Deze extra functionaliteiten zorgen jammer genoeg ook voor grote uitdagingen bij het behoud van spreadsheets.

Bronnen:
http://www.digitalpreservation.gov/formats/fdd/fdd000398.shtml (over .xlsx)
https://en.wikipedia.org/wiki/Microsoft_Excel
https://www.openoffice.org/sc/excelfileformat.pdf

10.2 Risico-inventarisatie

10.2.1 Verschillen in interpretatie na migratie

Het hangt er een beetje vanaf hoe spreadsheets worden gebruikt maar vaak bevat minimaal één cel een formule. Het probleem met deze formules is dat ze vaak software-afhankelijk zijn. Microsoft Office Excel interpreteert de formules soms anders dan bijvoorbeeld OpenOffice Calc. Hier moeten we rekening mee houden bij de preservering, bijv. als Excel-bestanden naar OpenOffice-bestanden omgezet zouden worden.

10.2.2 Verschillen in vormgeving na migratie

Ook kan de vormgeving van werkbladen bij het omzetten naar andere formaten veranderen of verloren gaan. Microsoft legt zelf uit wat er mogelijk verloren gaat bij de opslag (vanuit Excel) in een ander formaat: https://support.office.com/en-us/article/Excel-formatting-and-features-that-are-not-transferred-to-other-file-formats-8fdd91a3-792e-4aef-a5bb-46f603d0e585?ui=en-US&rs=en-US&ad=US.

10.2.3 Mate van openheid

Excelbestanden van voor 1997 vallen in de categorie "gesloten proprietair formaat". De versies van Excel 1997 tot 2010 zijn te classificeren als "open proprietair", wat een verminderd preservationrisico inhoudt. Sinds Microsoft Office Open XML (OOXML, .xlsx voor Excelbestanden) uitkwam is het preservationrisico nog lager.

10.2.4 Macro's

In Excel kan gebruik worden gemaakt van macro's. Hiervoor gebruik je Visual Basic for Applications (VBA), een variant van Visual Basic. Ondersteuning van VBA in andere spreadsheetformaten is minstens onvolledig. Zo heeft Apache OpenOffice de macrotaal OpenOffice Basic, "*a programming language similar to Microsoft Visual Basic for Applications (VBA). Apache OpenOffice has some Microsoft VBA macro support.*

OpenOffice Basic is available in Writer, Calc and Base." (https://en.wikipedia.org/wiki/Apache_OpenOffice) Bij Excelbestanden met veel, lange en/of complexe VBA-macro's gaat bij migratie naar andere formaten hoogstwaarschijnlijk informatie (en functionaliteit) verloren.

10.3 Evaluatie

Hoe goed gedocumenteerd, gestandaardiseerd en vrij van licenties/patenten ook, het Excel-formaat (incl. OOXML) is een formaat dat eigendom is van of alleen bewerkt kan worden door Microsoft, en daarmee geen volledig open formaat. De meest recente OOXML-versies vormen een beperkter preservationrisico.

Als gekozen wordt voor migratie van Excel naar andere spreadsheetformaten, dan moet goed onderzocht worden welke functionaliteiten bewaard moeten worden. Migratie van macro's is bijv. minstens onvolledig en ook kan bij migratie inhoud of layout verloren gaan.

10.4 Alternatieven

Zie voor overzichten van alternatieve spreadsheetsoftware:
- https://en.wikipedia.org/wiki/List_of_spreadsheet_software
- http://alternativeto.net/software/microsoft-excel/

Als het alleen om het bewaren van niet-interactieve informatie uit cellen gaat, kan het kommagescheiden (.csv) tekstbestand als alternatief voor een spreadsheet worden gekozen.

De meest in het oog springende open alternatieven die ook interactieve functies bieden lijken momenteel LibreOffice Calc (.ods, OpenDocument-formaat), Apache OpenOffice Calc (.ods, OpenDocument-formaat) en Gnumeric (voor Linux, .gnm of .gnumeric, een ge-gzipt XML-formaat).

Andere veelgebruikte alternatieven zijn proprietaire en/of gesloten formaten, zoals Google Drive Sheets (gesloten, proprietair en grotendeels ongedocumenteerd formaat) en Apple iWorks Numbers (gesloten, proprietair en grotendeels ongedocumenteerd formaat).

10.5 Ondersteuning in het e-Depot

10.5.1 Formaten

Zoeken naar "Microsoft Excel" in de registry levert 34 verschillende formaten op:

PUID	Name	Version
fmt/55	Microsoft Excel 2.x Worksheet (xls)	2
fmt/62	Microsoft Excel 2000-2003 Workbook (xls)	8X
fmt/56	Microsoft Excel 3.0 Worksheet (xls)	3
fmt/58	Microsoft Excel 4.0 Workbook (xls)	4W
fmt/57	Microsoft Excel 4.0 Worksheet (xls)	4S
fmt/59	Microsoft Excel 5.0/95 Workbook (xls)	mei-95
fmt/60	Microsoft Excel 95 Workbook (xls)	7
fmt/61	Microsoft Excel 97 Workbook (xls)	8
x-fmt/124	Microsoft Excel Add-In	
x-fmt/23	Microsoft Excel Backup	
fmt/553	Microsoft Excel Chart	2.x
fmt/554	Microsoft Excel Chart	3.0
x-fmt/126	Microsoft Excel Chart	4.0

PUID	Name	Version
fmt/555	Microsoft Excel Macro	2.x
fmt/556	Microsoft Excel Macro	3.0
x-fmt/123	Microsoft Excel Macro	4.0
fmt/445	Microsoft Excel Macro-Enabled	2007
fmt/595	Microsoft Excel Non-XML Binary Workbook	2007 onwards
x-fmt/46	Microsoft Excel ODBC Query	
x-fmt/74	Microsoft Excel OLAP Query	
x-fmt/97	Microsoft Excel OLE DB Query	
fmt/598	Microsoft Excel Template	2007 onwards
x-fmt/17	Microsoft Excel Template	2000
x-fmt/125	Microsoft Excel Toolbar	
x-fmt/58	Microsoft Excel Web Query	
x-fmt/128	Microsoft Excel Workspace	
fmt/172	Microsoft Excel for Macintosh	3.0
fmt/173	Microsoft Excel for Macintosh	4.0
fmt/174	Microsoft Excel for Macintosh	98
fmt/175	Microsoft Excel for Macintosh	2001
fmt/176	Microsoft Excel for Macintosh	2002
fmt/177	Microsoft Excel for Macintosh	2004
fmt/178	Microsoft Excel for Macintosh	X
fmt/214	Microsoft Excel for Windows	2007 onwards

10.5.2 Migration pathways

Het e-Depot kent diverse migration pathways om verschillende versies van MS Excelbestanden om te zetten naar een ander formaat:

PUID	Name	Version
pth/174	Open Office Convert XLS 2.1 to ODS 1.2	
pth/136	Open Office Convert XLS 2.1 to PDF 1.4	
pth/318	Open Office Convert XLS 2.1 to PDF/A	
pth/181	Open Office Convert XLS 2000-2003 to ODS 1.2	
pth/143	Open Office Convert XLS 2000-2003 to PDF 1.4	
pth/325	Open Office Convert XLS 2000-2003 to PDF/A	
pth/175	Open Office Convert XLS 3.0 to ODS 1.2	
pth/137	Open Office Convert XLS 3.0 to PDF 1.4	
pth/319	Open Office Convert XLS 3.0 to PDF/A	
pth/176	Open Office Convert XLS 4.0S to ODS 1.2	
pth/138	Open Office Convert XLS 4.0S to PDF 1.4	
pth/320	Open Office Convert XLS 4.0S to PDF/A	
pth/177	Open Office Convert XLS 4.0W to ODS 1.2	
pth/139	Open Office Convert XLS 4.0W to PDF 1.4	
pth/321	Open Office Convert XLS 4.0W to PDF/A	
pth/178	Open Office Convert XLS 5.0 to ODS 1.2	
pth/140	Open Office Convert XLS 5.0 to PDF 1.4	
pth/322	Open Office Convert XLS 5.0 to PDF/A	
pth/179	Open Office Convert XLS 95 to ODS 1.2	
pth/141	Open Office Convert XLS 95 to PDF 1.4	
pth/323	Open Office Convert XLS 95 to PDF/A	
pth/180	Open Office Convert XLS 97 to ODS 1.2	
pth/142	Open Office Convert XLS 97 to PDF 1.4	
pth/324	Open Office Convert XLS 97 to PDF/A	
pth/585	Open Office Convert XLS for Mac 2001 to ODS 1.2	
pth/561	Open Office Convert XLS for Mac 2001 to PDF 1.4	
pth/576	Open Office Convert XLS for Mac 2001 to PDF 1.4/A	
pth/586	Open Office Convert XLS for Mac 2002 to ODS 1.2	
pth/562	Open Office Convert XLS for Mac 2002 to PDF 1.4	
pth/577	Open Office Convert XLS for Mac 2002 to PDF 1.4/A	
pth/587	Open Office Convert XLS for Mac 2004 to ODS 1.2	

PUID	Name	Version
pth/563	Open Office Convert XLS for Mac 2004 to PDF 1.4	
pth/578	Open Office Convert XLS for Mac 2004 to PDF 1.4/A	
pth/582	Open Office Convert XLS for Mac 3.0 to ODS 1.2	
pth/558	Open Office Convert XLS for Mac 3.0 to PDF 1.4	
pth/573	Open Office Convert XLS for Mac 3.0 to PDF 1.4/A	
pth/583	Open Office Convert XLS for Mac 4.0 to ODS 1.2	
pth/559	Open Office Convert XLS for Mac 4.0 to PDF 1.4	
pth/574	Open Office Convert XLS for Mac 4.0 to PDF 1.4/A	
pth/584	Open Office Convert XLS for Mac 98 to ODS 1.2	
pth/560	Open Office Convert XLS for Mac 98 to PDF 1.4	
pth/575	Open Office Convert XLS for Mac 98 to PDF 1.4/A	
pth/588	Open Office Convert XLS for Mac v.X to ODS 1.2	
pth/564	Open Office Convert XLS for Mac v.X to PDF 1.4	
pth/579	Open Office Convert XLS for Mac v.X to PDF 1.4/A	
pth/380	Open Office Convert XLSX to ODS 1.2	
pth/375	Open Office Convert XLSX to PDF 1.4	
pth/378	Open Office Convert XLSX to PDF/A	

10.5.3 Software en Tools

Beschikbare software voor het omgaan met MS Excel bestaat uit:

PUID	Name	Version
x-sfw/20	Excel	2000 (9.0) for Windows
x-sfw/21	Excel	97 (8.0) for Windows
x-sfw/22	Excel	95 (7.0) for Windows
x-sfw/23	Excel	5.0
x-sfw/24	Excel	4.0
x-sfw/25	Excel	3.0
x-sfw/26	Excel	2.1
x-sfw/270	Excel	XP
x-sfw/276	Excel	2003

Er zijn in Preservica geen tools beschikbaar specifiek voor Excel (.xls en .xlsx).

10.6 Voorgestelde strategie

De voorgestelde strategie voor Excelbestanden is dat momenteel geen acties noodzakelijk zijn, omdat sommige Excelformaten weliswaar verouderd, maar nog altijd leesbaar zijn en voldoende door software ondersteund worden.

Periodiek moet nagegaan worden of deze strategie aangepast moet worden. Ook omdat Microsoft Office Excel marktleider is op het gebied van spreadsheets, alternatieve spreadsheetproducten (met een open formaat) soms minder ver uitontwikkeld zijn, en er informatieverlies kan optreden bij migratie naar andere formaten, zijn er meerdere mogelijkheden:

- als alleen de data (en niet de interactiviteit van eventuele formules en/of layout) van belang zijn: migratie van Excel-bestanden naar kommagescheiden (.csv) tekstbestanden
- als naast de data ook (enige) functionaliteit en/of layout bewaard moet blijven: migratie van Excel-bestanden naar het OpenDocument-formaat voor spreadsheets (.ods)
    - Let op: volledig behoud van informatie (data, functionaliteit en layout) is verre van verzekerd en de kwaliteit moet van geval tot geval gecontroleerd worden!

- als volledig behoud van informatie (data, functionaliteit en layout) noodzakelijk is en de ondersteuning van het (ver)oude(rde) Excel-formaat onvolledig is of stopt: periodieke conversie van (ver)oude(rde) Excel-bestanden naar de meest recente versie, momenteel OOXML

NB: voor *presentatie*doeleinden kan eventueel gekozen worden voor het beschikbaarstellen van een PDF- of HTML-versie.

In alle gevallen heeft het de voorkeur het originele Excel-bestand te bewaren, omdat er in de toekomst nieuwe mogelijkheden voor preservering kunnen ontstaan.

11 MS Powerpoint

11.1 Algemene informatie

Microsoft PowerPoint is een is een pakket waarmee vooral gemakkelijk presentaties gemaakt kunnen worden die bestaan uit meerdere dia's. Door het gebruik van een eenduidige opmaak kan men eenvoudig mooie, eenduidige presentaties maken. Ook kan er gesproken woord en muziek worden ingevoegd. Door gebruik van hyperlinks kan men per dia verwijzen naar bijvoorbeeld andere presentaties en filmpjes. Sinds de PowerPoint versie 2003 is het standaard mogelijk een presentatie zelfstartend op een zelfstandige drager (bijvoorbeeld een cd of USB stick) te zetten inclusief een viewer. Daardoor kan de presentatie op elke computer worden afgespeeld, zonder dat PowerPoint geïnstalleerd hoeft te zijn. Powerpoint maakt deel uit van de Microsoft Office suite.

Geschiedenis:
- April 1987: PowerPoint 1, door Forethought, Californië.
- Augustus 1987: Microsoft koopt het bedrijf Forethought.
- Mei 1988: PowerPoint 2
- Mei 1990: PowerPoint 2 voor Windows
- Mei 1992: PowerPoint 3 voor Windows 3.1
- PowerPoint 3 voor Windows.
- PowerPoint 4, 7, 97, 2000, XP, 2003, 2007, 2010, 2013: onderdeel van Microsoft Office.

11.2 Risico-inventarisatie

11.2.1 Gesloten (proprietair) formaat

Het Powerpoint 97-2002 formaat is het eigen ("native") formaat van Powerpoint 97 en latere versies. Het formaat is proprietair en Microsoft maakt de details van de structuur niet openbaar. Powerpoint 97-2002 formaat gebaseerd op het generieke OLE2 Compound Document Format. Een Powerpoint presentatie wordt opgeslagen als een PowerPoint Document stream binnen een Compound Document Format file. Het formaat bleef onveranderd met de release van Powerpoint 2000 en XP.[37]

Zoals gezegd is vanaf Microsoft Office 2007 het standaard output formaat van MS Powerpoint gebaseerd op het meer open OOXML file formaat. [38]

11.2.2 Multimedia content [39]

Powerpoints kunnen multimedia content bevatten: filmpjes, geluidsfragmenten en 3D content. Rendering van multimedia content kan afhankelijk zijn van externe applicaties die in de toekomst misschien niet beschikbaar zijn; daarnaast kan het formaat waarin de multimedia content is opgeslagen verouderd raken.

[37] Bron: Preservica 5.4 registry https://e-Depot-acpt.nationaalarchief.nl/Registry/registry.html#list:fmt/126&false (geraadpleegd 16-11-2015)
[38] Bron: Preservica 5.4 registry https://e-Depot-acpt.nationaalarchief.nl/Registry/registry.html#list:fmt/215&false (geraadpleegd 16-11-2015)
[39] Bron: Afgeleid van PDF-tests: http://wiki.opf-labs.org/display/TR/Multimedia+content (geraadpleegd 22-10-2015)

11.3 Evaluatie

Hoe goed gedocumenteerd, gestandaardiseerd en vrij van licenties/patenten ook, het Powerpoint-formaat (incl. OOXML) is een formaat dat eigendom is van of alleen bewerkt kan worden door Microsoft, en daarmee geen volledig open formaat. De meest recente OOXML-versies vormen een beperkter preservationrisico.
Als gekozen wordt voor migratie van Powerpoint naar andere presentatieformaten, dan moet goed onderzocht worden welke functionaliteiten bewaard moeten worden.

11.4 Alternatieven

Zie voor overzichten van alternatieve presentatie software:
- https://en.wikipedia.org/wiki/Category:Presentation_software
- http://alternativeto.net/software/microsoft-powerpoint/

De meest in het oog springende open alternatieven die ook interactieve functies bieden lijken momenteel LibreOffice Impress (.odp, OpenDocument-formaat) en Apache OpenOffice Impress (.odp, OpenDocument-formaat).

Andere veelgebruikte alternatieven zijn proprietaire en/of gesloten formaten, zoals Google Drive Slides (gesloten, proprietair en grotendeels ongedocumenteerd formaat) en Apple iWorks Keynote (gesloten, proprietair en grotendeels ongedocumenteerd formaat).

11.5 Ondersteuning in het e-Depot

11.5.1 Formaten

De registry bevat geen formaten die voldoen aan de zoekterm "ppt" of "pptx". De zoekterm "powerpoint" levert wel 14 treffers op.

PUID	Name	Version
fmt/487	Macro Enabled Microsoft Powerpoint	2007 Onwards
x-fmt/177	Microsoft PowerPoint Graphics File	
fmt/179	Microsoft PowerPoint for Macintosh	4.0
fmt/180	Microsoft PowerPoint for Macintosh	98
fmt/181	Microsoft PowerPoint for Macintosh	2001
fmt/182	Microsoft PowerPoint for Macintosh	X
x-fmt/86	Microsoft Powerpoint Add-In	
x-fmt/84	Microsoft Powerpoint Design Template	
x-fmt/216	Microsoft Powerpoint Packaged Presentation	
fmt/125	Microsoft Powerpoint Presentation	95
fmt/126	Microsoft Powerpoint Presentation	97-2002
x-fmt/88	Microsoft Powerpoint Presentation	4.0
x-fmt/87	Microsoft Powerpoint Show	
fmt/215	Microsoft Powerpoint for Windows	2007 onwards

11.5.2 Migration pathways

Er zijn zes migration pathways om Powerpoint bestanden om te zetten naar een ander formaat: drie voor .ppt en 3 voor .pptx bestanden.

PUID	Name	Version
pth/182	Open Office Convert PPT 97-2002 to ODP 1.2	
pth/145	Open Office Convert PPT 97-2002 to PDF 1.4	
pth/329	Open Office Convert PPT 97-2002 to PDF/A	
pth/381	Open Office Convert PPTX to ODP 1.2	
pth/374	Open Office Convert PPTX to PDF 1.4	
pth/377	Open Office Convert PPTX to PDF/A	

11.5.3 Software en Tools

Beschikbare software voor het omgaan met .ppt en .pptx bestaat uit:

PUID	Name	Version
x-sfw/278	Powerpoint	XP
x-sfw/40	Powerpoint	2000 (9.0)

Er zijn in Preservica geen tools beschikbaar specifiek voor Powerpoint (.ppt en .pptx).

11.6 Voorgestelde strategie

Uitgaande van het just in time principe is er op moment van schrijven geen actie nodig. Op dit moment zijn er geen grote preservation risico's met Powerpointbestanden. Wel is het raadzaam om dit periodiek opnieuw te beoordelen (d.m.v. preservation watch). Als er zich inderdaad risico's voordoen met Powerpoint zijn er meerdere alternatieven te bedenken.

Ook omdat Microsoft Office Powerpoint marktleider is op het gebied van presentatiesoftware, alternatieven (met een open formaat) soms minder ver uitontwikkeld zijn, en er informatieverlies kan optreden bij migratie naar andere formaten, zijn er meerdere mogelijkheden:

- migratie van powerpoint-bestanden naar PDF/A bestanden
- migratie van powerpoint-bestanden naar het OpenDocument-formaat voor presentaies (.odp)
  Let op: volledig behoud van informatie (data, functionaliteit en layout) is verre van verzekerd en de kwaliteit moet van geval tot geval gecontroleerd worden!
- als volledig behoud van informatie (data, functionaliteit en layout) noodzakelijk is en de ondersteuning van het (ver)oude(rde) Powerpoint-formaat onvolledig is of stopt: periodieke conversie van (ver)oude(rde) Powerpoint-bestanden naar de meest recente versie, momenteel OOXML

In alle gevallen heeft het de voorkeur het originele Powerpoint-bestand te bewaren, omdat er in de toekomst nieuwe mogelijkheden voor preservering kunnen ontstaan.

12 MS Access

12.1 Algemene informatie

Wikipedia meldt dat een database, gegevensbank of databank een digitaal opgeslagen archief is, ingericht met het oog op flexibele raadpleging en gebruik. Databases spelen een belangrijke rol voor het archiveren en actueel houden van gegevens bij onder meer de overheid, financiële instellingen en bedrijven, in de wetenschap, en worden op kleinere schaal ook privé gebruikt.

Microsoft Office Access is een database management systeem (DBMS) van Microsoft, bestemd voor gebruik op desktopcomputers. (Het product Microsoft SQL Server is bestemd voor client-servergebruik, en Access kan desgewenst naar SQL Server worden geüpgraded. Microsoft noemt dit upsizen.) Access bestaat uit de relational Microsoft Jet Database Engine, een grafische gebruikersinterface en softwareontwikkeltools, en is onderdeel van Microsoft Office. Accessbestanden maken geen gebruik van OOXML en de algemene uitgangspunten van MS Office.

Volgens Wikipedia waren Borland (met Paradox en dBase) en Fox (met FoxPro) ooit de dominante spelers op de desktopdatabasemarkt. Microsoft Access was het eerste Windows-databaseprogramma voor het grote publiek en groeide in de jaren ’90 uit naar de positie van dominante speler voor Windows-gebaseerde databases, mede omdat veel andere spelers de overstap van MS-DOS naar Windows niet snel of goed genoeg maakten.

Access slaat data op in een eigen formaat, gebaseerd op de Access Jet Database Engine, en kan data uit andere (database)systemen importeren of er mee linken. Sinds Access 2007 is het standaard bestandsformaat voor Access-databases .accdb. Daarvoor was het .mdb.

Access kan gebruik maken van Visual Basic for Applications (VBA), en via VBA bijvoorbeeld gebruik maken van ActiveX-componenten en functies van het Windows-besturingssysteem.

Meer informatie over databases in het algemeen en Microsoft Office Access in het bijzonder:

- Databases
    - https://nl.wikipedia.org/wiki/Database
    - http://openpreservation.org/system/files/Database%20archiving%20review.pdf
- Access
    - https://en.wikipedia.org/wiki/Microsoft_Access
    - https://en.wikipedia.org/wiki/Microsoft_Jet_Database_Engine
    - http://fileformats.archiveteam.org/wiki/Access
    - https://support.office.com/en-ZA/article/which-access-file-format-should-i-use-012d9ab3-d14c-479e-b617-be66f9070b41

12.2 Risico-inventarisatie

12.2.1 Mate van openheid

De Access-formaten zijn gesloten, proprietair en de specificaties zijn niet gepubliceerd. Via reverse engineering is het bestandsformaat voor Jet3 en Jet4-databases echter toch redelijkerwijs gedocumenteerd, zie bijv.: https://github.com/brianb/mdbtools/blob/master/HACKING. Access is daarmee geen open, valideerbaar en volledig gedocumenteerd bestandsformaat dat voldoet aan een open standaard.

12.2.2 Niet terugwaarts compatibel

Access is ook niet volledig backward compatible. Op diverse momenten in de ontwikkelcyclus van Access is de terugwaartse compatibiliteit gebroken. Wikipedia zegt hier bijv. over dat: "*[the] native Access database format (the Jet MDB Database) has also evolved over the years. Formats include Access 1.0, 1.1, 2.0, 7.0, 97, 2000, 2002, 2007, and 2010. The most significant transition was from the Access 97 to the Access 2000 format; which is not backward compatible with earlier versions of Access. As of 2011 all newer versions of Access support the Access 2000 format. New features were added to the Access 2002 format which can be used by Access 2002, 2003, 2007, and 2010.*"

12.3 Evaluatie

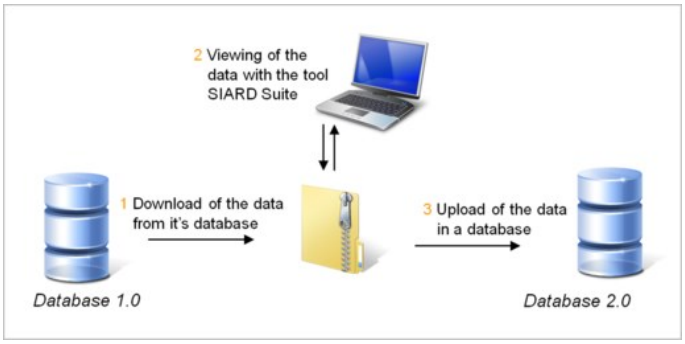
Microsoft Office Access (zowel .accdb als .mdb) is een gesloten, proprietair en niet openbaar gedocumenteerde industry standard voor desktopdatabases. Het voldoet daarmee niet aan de overbrengingsvereisten van de Archiefwet.

Omdat Microsoft met Access echter wereldwijd marktleider is op het gebied van desktopdatabases, wordt Access veel gebruikt en moeten we er als NA rekening mee houden dat we regelmatig met Access-databases in aanraking zullen komen. Volgens onze algemene uitgangspunten verdient het dan de voorkeur niet zomaar een Accessdatabase te converteren naar een andere, open standaard. Hierdoor zou nl. informatieverlies kunnen optreden. Het is beter vooraf te overleggen over de mogelijkheden.

12.4 Alternatieven

Voor Microsoft Office Access als DBMS zijn diverse alternatieven. Hier beperken we ons tot die alternatieven die voldoen aan de criteria van een open, valideerbaar en volledig gedocumenteerd bestandsformaat dat voldoet aan een open standaard, en/of die alternatieven die gangbaar zijn voor de preservation van databases.

SIARD (https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html) staat voor Software Independent Archiving of Relational Databases en is ontwikkeld door de Swiss Federal Archives om relationele databases duurzaam te kunnen archiveren. SIARD bestaat uit een formaat en een suite. Het SIARD-formaat is een open formaat voor het archiveren van de inhoud van relationele databases. De SIARD-suite is een freewaretool voor het omzetten van relationele databases (momenteel: Oracle, Microsoft SQL Server, MySQL, DB/2 en Microsoft Access) naar het SIARD-formaat.Op het moment van schrijven wordt in het kader van het Europese EARK-project gewerkt aan een nieuwe versie van de SIARD-suite en het SIARD-formaat.



Figuur 2 - Archiving with SIARD Suite

Een ander alternatief voor Microsoft Access is de Database Markup Language (DBML), een in het kader van het RODA-project ontwikkeld XML-formaat voor de opslag van de structuur en de inhoud van relationele databases. (http://conferences.idealliance.org/extreme/html/2007/Ramalho01/EML2007Ramalho01.html)

Er is geen officiële open standaard voor het archiveren van databases. SIARD en DBML komen als de facto standaard uit de literatuur naar voren, en SIARD lijkt de best ondersteunde en doorontwikkelde van de twee te zijn, momenteel bijvoorbeeld in het kader van het EC-project E-ARK.

NB: het door DANS/MIXED ontwikkelde Standard Data Format for Preservation (SDFP) is een tussenformaat gebaseerd op SIARD voor databases en ODF voor spreadsheets (https://sites.google.com/a/datanetworkservice.nl/mixed/documentation). Dhr. Dirk Roorda van DANS gaf in november 2015 aan dat SFDP "*het niet heeft gehaald*". We laten SDFP daarom verder buiten beschouwing.

De db-preservation-toolkit (http://www.database-preservation.com/) is een nuttig middel om databases (incl. Microsoft Access) van het ene formaat naar het andere te converteren. De toolkit is via een LGPL-licentie verkrijgbaar op GitHub, waar hij voornamelijk wordt onderhouden door KEEP Solutions: https://github.com/keeps/db-preservation-toolkit. Momenteel wordt de toolkit, samen met een nieuwe versie van het SIARD-formaat, doorontwikkeld in het kader van het Europese project E-ARK.

12.5 Ondersteuning in het e-Depot

Omdat het e-Depot migratie als strategie heeft geïmplementeerd, leent het zich voornamelijk voor het bewaren van bevroren exports of snapshots van databases. Als databases nog benaderd of geëmuleerd moeten kunnen worden, dan moet dat buiten en los van het e-Depot gerealiseerd worden.

12.5.1 Formaten

De registry bevat 6 formaten die voldoen aan de zoekterm "Access":

fmt/275	Microsoft Access Database	2007
x-fmt/238	Microsoft Access Database	95
x-fmt/239	Microsoft Access Database	97
x-fmt/240	Microsoft Access Database	2000
x-fmt/241	Microsoft Access Database	2002
x-fmt/66	Microsoft Access Database	2.0

12.5.2 Migration pathways

Er zijn geen migration pathways om Accessbestanden om te zetten naar een ander formaat. Dit zou buiten het e-Depot moeten worden gerealiseerd, of als functionaliteit aan het e-Depot moeten worden toegevoegd.

12.5.3 Software en Tools

Beschikbare software voor het omgaan met Accessbestanden bestaat uit:

PUID	Name	Version
x-sfw/41	Access	2000 (9.0)

Er zijn in het e-Depot geen tools beschikbaar specifiek voor Access.

12.6 Voorgestelde strategie

Voor het duurzaam bewaren van (Access)databases zijn meerdere mogelijkheden.

12.6.1 Minder dan 10 jaar

Voor (Access)databases met een korte bewaartermijn (korter dan 10 jaar) is migratie in de vorm van achterwaartse compatibiliteit een goede bewaarstrategie om databases inclusief de gebruikersapplicatie tijdens hun beperkte levensloop met behoud van authenticiteit en integriteit toegankelijk te houden. Ofschoon het mogelijk is databases te bewaren in het oorspronkelijke bestandsformaat, heeft het de voorkeur de databases te upgraden naar het nieuwere bestandsformaat, aangezien software slechts een beperkt aantal generaties oudere bestandsformaten betrouwbaar kan lezen. Zo zouden .mdb-bestanden geüpgraded kunnen worden naar .accdb-bestanden[40]. Steekproefsgewijze, visuele controle is altijd noodzakelijk om vast te stellen of de migratie tot het gewenste resultaat heeft geleid. Met andere woorden: wordt er nog voldaan aan de door de organisatie gestelde authenticiteitseisen.

12.6.2 Langer dan 10 jaar

De omzetting van databases naar XML is een bewaarstrategie die in staat is om databases die voor de lange termijn bewaard moeten blijven (langer dan 10 jaar) in authentieke staat te representeren. De meest geschikte optie voor deze migratie is SIARD.

12.6.3 Het origineel bewaren

Het is aan te raden om ook het originele bestand te bewaren, om maximale flexibiliteit te bieden voor toekomstige bewaarstrategieën. Zolang het originele bestandsformaat nog breed toegankelijk is, biedt dit tevens de 'meest authentieke' representatie van de database, met name in combinatie met een nog werkende gebruikersapplicatie.

[40] Zie https://support.office.com/en-us/article/Convert-a-database-to-the-accdb-file-format-69abbf06-8401-4cf3-b950-f790fa9f359c (geraadpleegd op 7-12-2015)