

Bestandstrategie opname bestandsformaten e-depot

Inhoud

[Inleiding 3](#_Toc44588101)

[Bestandsformaten 4](#_Toc44588102)

[PDF / ODT Document 6](#_Toc44588103)

[Office document 8](#_Toc44588104)

[Spreadsheet 9](#_Toc44588105)

[E-mail 10](#_Toc44588106)

[Afbeelding 11](#_Toc44588107)

[Presentatie 12](#_Toc44588108)

[Databank 13](#_Toc44588109)

[Vector Image 14](#_Toc44588110)

[Audio 15](#_Toc44588111)

[Video 16](#_Toc44588112)

[Website 17](#_Toc44588113)

# Inleiding

Preservering van informatieobjecten kan op verschillende manieren uitgevoerd worden. Vooraf bij selectie en opname af achteraf indien een bestandsformaat een bedreiging gaat vormen. Door de continue ontwikkelingen op het gebied van informatie technologie kan het voorkomen dat een bestandsformaat op een bepaald moment onleesbaar wordt. Bijvoorbeeld indien de software waarmee het bestand uitgelezen kan worden niet meer beschikbaar is. Voordat een dergelijke bedreiging ontstaat kan door achteraf te preserveren of gebruik te maken van een toekomstige oplossingen zoals bijvoorbeeld emulatie het bestand alsnog toegankelijk gemaakt worden.

Het West-Brabants Archief (WBA) heeft een werkgroep opgezet om het onderdeel preservering verder uit te werken en te integreren in het proces van opname in het e-depot. Als uitgangspunt hiervoor is gekozen om dit zoveel mogelijk conform het preserveringsbeleid van het Nationaal Archief te gaan doen. Doordat toekomstige oplossingen mogelijk meer kunnen bieden dan de huidige preservering-technieken is gekozen voor preservering enkel bij opname met daarbij zoveel mogelijk behoud van het originele bestand. Om te zorgen voor een goede bewaring en toegankelijkheid van integere en authentieke informatieobjecten worden enkel bestandsformaten toegelaten tot het e-depot welke dit voor de langere termijn kunnen garanderen. In het document bestandstrategieën van het Nationaal archief is een overzicht van voorkeurs- en acceptabele bestandsformaten te vinden. De reden om enkel deze formaten toe te staan is om dataverlies en dubbele opslag bij conversie door directe preservering te voorkomen. Uit steekproeven bij de aangesloten gemeenten is ook gebleken dat het grootste gedeelte van de huidige digital born informatieobjecten voorkomen op de lijst van de toegestane formaten. Indien een bestandsformaat aangetroffen wordt welke niet op de lijst voorkomt zal het WBA met de informatieproducent in gesprek gaan. Hierbij wordt dan gekeken of een afwijkend formaat zonder te veel verlies naar één van de voorkeurs- of acceptabele formaten kan worden geconverteerd vóór opname in het e-depot bij de informatieproducent zelf. Indien dit niet het geval is kan gekozen worden voor een uitzondering waarbij het origineel toch opgenomen wordt in het e-depot om dataverlies te voorkomen. Toekomstige technologische ontwikkelingen kunnen hiervoor alsnog een oplossing bieden. Deze strategie zal in de komende jaren verder onderzocht worden om hiermee het preserveringsbeleid door te ontwikkelen.

# Bestandsformaten

Het WBA heeft op basis van de lijst voorkeurs- en acceptabele formaten een selectie gemaakt om van elk een bestand op te nemen in het e-depot. Standaard bevat het e-depot allerlei preserveringregels welke ge(de)activeerd of aangepast kunnen worden. De ervaringen tijdens het doorlopen van deze procedure heeft geleid tot een overzicht van formaten waarbij gekeken is of door het gebruik van open formaten of veel voorkomende (niet open) formaten te verwachten valt dat deze voor minimaal 5 tot 10 jaar geen risico zullen vormen in een beperking van de toegang.

|  |  |
| --- | --- |
| **Type** | **Bestandformaat** |
| Document | PDF/a – ODT – DOC – DOCX |
| Spreadsheet | ODS – PDF/a – CSV – XLS – XLSX |
| E-mail | EML – MSG – PST – MBOX |
| Afbeelding | TIFF – PNG – JPEG – JP2 – JPX |
| Presentatie | ODP – PDF/a – PPT – PPTX |
| Databank | XML |
| Vector | SVG – EPS |
| Audio | WAV – MP3 – FLAC |
| Video | MXF – MPEG/4 – MKV |
| Website | WARC |

Van de bovenstaande formaten is bekeken welke versie wenselijk is en welke niet. Ieder formaat kan namelijk meerdere versies bevatten met daarin soms grote verschillen in de syntaxis en semantiek waarop de informatie vastgelegd wordt. Dit heeft te maken met de doorontwikkeling van een bestandsformaat. Zo stamt het eerste MS Word document (DOC) uit begin jaren 80. Een .doc formaat uit bijvoorbeeld 1983 gemaakt onder MS DOS heeft een andere structuur en codering dan een .doc uit MS Office 2003 vele jaren later. Om risico’s met oudere bestandsversies te voorkomen en doordat de kans zeer klein is dat van een digital born document ouder dan 20 jaar geen papieren versie is, hebben wij gekozen om enkel versies jonger dan 20 jaar toe te staan. Ook al zijn er wel informatieobjecten bij de informatieproducent te vinden van bijvoorbeeld voor 2000 deze zijn dan ook altijd op papier aanwezig. Archieven worden geheel als papier overgebracht indien papier leidend is en dus worden deze oudere digitale bestanden niet overgebracht naar het e-depot. De ervaring bij de aangesloten gemeenten in West-Brabant is dat de eerste volledig digitale archieven pas na circa 2010 gevormd zijn.

Hieronder volgt een overzicht per documenttype met daarin van ieder bestandsformaat welke versie geaccepteerd wordt en de bevindingen en/of nog openstaande vragen. Dit zijn de regels welke op dit moment gehandhaafd worden. Zeer waarschijnlijk zullen de onderstaande resultaten en de lijst van geaccepteerde formaten de komende jaren worden bijgesteld. Ook container bestanden en formaten van vak specifieke applicaties zullen we tegen gaan komen, deze zijn vaak moeilijk of enkel met verlies te converteren. Kortom de ontwikkelingen op dit gebied van preservering zullen de komende jaren zeker niet stil blijven staan.

## PDF / ODT Document

Alle PDF/A formaten met uitzondering van PDF/A-3 maken dat het primaire document voldoende beschermd is tegen preserveringsrisico’s op de (zeer) lange termijn. Voor PDF/A-3 gebruik geldt dat er voor archivering een workflow wordt gedefinieerd die garandeert dat de relatie tussen het pdf-document en het ingesloten bestand volkomen duidelijk is voor het WBA. Dit vereist dus vastgelegde afspraken tussen de informatieproducenten en het WBA.

Conversie van PDF-bestanden naar PDF/A-formaat wordt afgeraden, omdat het foutgevoelig is en er zijn geen systemen bekend voor die foutherkenning erop.

Sommige applicaties produceren verminkte PDF-1.7 of oudere versies bestanden, welke niet volledig voldoen aan (de specificaties van) het PDF-formaat. Toekomstige migraties hiermee kunnen mislukken en leiden tot dataverlies. PDF is een complex formaat, waardoor herstel lastig is, mede door het gebrek aan goede hulpmiddelen daarvoor.

Extra lastig is tevens dat PDF het gebruik van encryptie toestaat. Dat betekent dat de toegang of hergebruik van de inhoud van een document wordt afgeschermd met een wachtwoord, of dat een bepaalde functionaliteit (denk aan afdrukken of tekst-onderdelen eruit kopiereen) wordt uitgezet. Deze restricties bemoeilijken toekomstige preserveringsacties of maken deze zelfs bij voorbaat zinloos.

Wanneer men deze PDF-soorten in Windows Verkenner ziet staan, zijn ze niet van elkaar te onderscheiden. Het is belangrijk deze drie typen wel te kunnen herkennen, namelijk op welke wijze ieder bestand tot stand gekomen is. Er bestaat een PDF validator tool om de oorsprong van elke pdf te achterhalen.

PDF/A is een wereldwijde open standaard, voorzien van NEN/ISO-norm. In het algemeen bestaan er echter risico’s mbt encryptie en onvolledige fronts binnen het PDF-formaat. Specifiek PDF/A-3 is echter niet aan te bevelen, gezien de hier boven genoemde risico’s. Hoewel er meerdere alternatieven voor PDF beschikbaar zijn, zoals LibreOffice, OpenOffice.org en AbiWord, zullen deze in de praktijk niet voor opslag in het e-depot gebruikt worden door het wereldwijde marktleiderschap van Adobe PDF.

PDF/A-1 en PDF/A-2 zijn als versies ISO 19005 opgenomen in de lijst van open standaarden van het Forum Standaardisatie en kunnen daarmee gezien worden als geaccepteerd.

Vanwege de foutgevoeligheid wensen wij geen conversies van PDF naar PDF/A.

Wij accepteren geen PDF-bestanden die gebruik maken van encryptie. En bestanden met ingepakte multimedia content adviseren wij die te extraheren en als apart bestand bij de pdf (zogenaamd supplementary file object) te ingesten. In overleg met het WBA kan een oplossing besproken worden.

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Document PDF** | **Opmerkingen** |
| **Opname** | PDF / ODT | PDF enkel acceptatie versies: 1.7 / 1.4: A-1a / A-2a / A-3a  **Op te nemen versies:** PDF: fmt/276, fmt/95, fmt/354, fmt/476, fmt/477, fmt/478, fmt/479, fmt/480, fmt/481  ODT: fmt/136, fmt/290, fmt/291 |
| **Preservering** | geen | Preserveren indien kwetsbaar (“Just in time”) |
| **Download** | origineel |  |
| **Opmerking** |  | Wat te doen met 1.7; Bijlagen en ingesloten objecten kunnen een risico (virus e.d.) vormen. Ontleden, niet accepteren of preserveren met mogelijk verlies. Is er een tool om 1.7 te onderzoeken op risico’s? Let op forum standaardisatie; mogelijk wordt 1.7 van de lijst duurzame formaten afgehaald (eind 2018). Mogelijke oplossing: PDF 1.7 preserveren naar PDF A-3 (ondersteund ook objecten).  PDF document: enkel opname multi-paged documenten, geen versnippering. Anders dient iedere pagina als bestand in de metadata beschreven te worden wat niet wenselijk is  Vanwege het risico tot dataverlies en/of corruptie van bestanden bij conversie wordt geadviseerd de originele van PDF versie 1.3 t/m 1.7 op te nemen en dus niet te converteren naar PDF a/1b. Indien nodig kunnen deze alsnog op termijn gepreserveerd worden naar een van de geaccepteerde formaten. |

## Office document

Word is een populaire applicatie, die voor teksten (brieven, nota’s, rapporten, etc.) wordt toegepast. Het bestandstype ervan kent vele vormen. Hoe goed gedocumenteerd, gestandaardiseerd en vrij van licenties of patenten ook, het Word-formaat is een bestandsvorm dat eigendom is van Microsoft en daarmee geen volledig open formaat. De meest recente OOXML-versies vormen een beperkter preservatie-risico. Er zijn inmiddels wel 20 verschillende formaten voor het Word bestandstype .DOC en er zijn voor het e-depot meerdere mogelijkheden om de verschillende versies van MS Word om te zetten naar een ander format (HTML, Portable Document Format 1,4, of XML 1.0).

De voorgestelde strategie voor Word bestanden is dat momenteel geen acties noodzakelijk zijn, omdat alle Word formaten nog leesbaar zijn en voldoende door software ondersteund worden.

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Document DOC** | **Opmerkingen** |
| **Oorspronkelijk** | DOC / DOCX | DOC acceptatie enkel vanaf doc versie 97 (ouder is vaak geen digital born en/of ook op papier archief)  **Op te nemen versies:** DOC: fmt/40, fmt/609  DOCX: fmt/412 |
| **Opname** | DOC / DOCX |  |
| **Preservering** | geen | Preserveren indien kwetsbaar (“Just in time”) |
| **Download** | origineel |  |
| **Opmerkingen** |  |  |

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 ontwikkeld zijn en er informatieverlies kan optreden bij migratie naar andere formaten, wordt er gekozen het originele bestandsformaat te behouden. Gezien de vele versies wordt gekozen om enkel de formaten van na 1997 op te nemen.

## Spreadsheet

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Document PDF** | **Opmerkingen** |
| **Oorspronkelijk** | ODS / PDF-A / CSV / XLS / XLSX | XLS accepteren vanaf 2000 (ivm ouder geen digital born of ook op papier)  **Op te nemen versies:** PDF-A: zie PDF  ODS: fmt/137, fmt/294, fmt/295  CSV: x-fmt/18  XLS: fmt/175, fmt/176, fmt/177, fmt/61, fmt/62  XLSX: fmt/214, fmt/494 |
| **Opname** | ODS / PDF-A / CSV / XLS / XLSX |  |
| **Preservering** | Geen | Geen preservering i.v.m. dataverlies. Alle formaten zijn voldoende open, enkel XLS niet. Echter door de lange levensduur niet direct een risico op de middellange termijn. |
| **Download** | origineel |  |
| **Opmerkingen** |  | PDF-A heeft niet de voorkeur voor een spreadsheet vanwege groot risico informatieverlies. Dit als advies meegeven aan de informatieproducent. Indien zowel PDF als origineel aangeleverd, enkel origineel accepteren.  Controleert het e-depot of RM tool op wachtwoord beveiligde spreadsheets? |

## E-mail

De protocollen en formaten voor het verzenden, transporteren en ontvangen van email zijn vrijwel allemaal gestandaardiseerd, open en goed gedocumenteerd. De grootste uitdagingen voor duurzame toegankelijkheid van email liggen aan de randen van het systeem: bij de opslag, de formaten waarin dat dient te gebeuren en de mate waarin het vervolgens kan worden geëxporteerd naar gestandaardiseerde, open formaten die duurzaam toegankelijk gehouden kunnen worden. Dankzij open en goed gedocumenteerde protocollen en standaarden voor email, zoals IMF, MIME, SMTP, IMAP en POP, is email goed te ondersteunen en te preserveren.

Voor opslag van email zijn diverse alternatieven, zoals opslag als PST, EML, XML, HTML, PDF of PDF/A. Hoewel er geen standaard bestaat, wordt overwegend .EML als opslagformaat toegepast.

Bij opname van email in het e-depot, worden de onderstaande formaten geaccepteerd. EML dient daarbij als voorkeursformaat, deze wordt nu namelijk gezien als een open standaard. Wel moet periodiek nagegaan worden in hoeverre er nieuwe emailopslag formaten en –strategieën opkomen, zoals migratie naar .XML of PDF(/A).

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Email** |  |
| **Opname** | EML / MSG / PST / MBOX | **Op te nemen versies:**  EML: fmt/278, fmt/950  MSG: x-fmt/430, fmt/1144  PST: fmt/248, fmt/249  MBOX: fmt/720 |
| **Preservering** | origineel | Indien geaccepteerd origineel behouden. |
| **Opmerkingen** |  | Email met bijlagen wordt geaccepteerd indien minimaal op documentniveau beschreven. Kan ook als één document met twee bestanden of twee documenten met ieder een bestand.  PST & MBOX is mogelijk een risico (vergelijkbaar met ZIP). Indien aanwezig vergt dit een onderzoek op locatie. Indien e-mails als project gearchiveerd worden kan een dergelijk container bestand een oplossing bieden. Eén dossier met daarbinnen losse e-mails heeft dan toch de voorkeur. |

## Afbeelding

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Afbeelding** | **Opmerkingen** |
| **Oorspronkelijk** | TIFF / PNG / JPEG / JP2 / JPX | **Op te nemen versies:** TIFF: fmt/153, fmt/154, fmt/155, fmt/156, fmt/353, x-fmt/387, x-fmt/388, x-fmt/399  PNG: fmt/11, fmt/12, fmt/13  JPEG: fmt/41, fmt/42, fmt/43, fmt/44  JP2: x-fmt/392 JPX: fmt/151 |
| **Opname** | TIFF / PNG / JPEG / JP2 / JPX |  |
| **Preservering** | TIFF / PNG / JPEG / JP2 / JPX | origineel behouden zonder preservering. Indien TIFF door het formaat moeilijk toegankelijk is dan preserveren naar Jpeg. |
| **Download** | TIFF / PNG / JPEG / JP2 / JPX |  |
| **Opmerking** |  | TIFF voorwaarden: on-gecomprimeerd (of lossless) enkel single paged images. |

## Presentatie

Microsoft PowerPoint is een pakket binnen de Microsoft Office Suite, waarmee gemakkelijk presentaties gemaakt kunnen worden die bestaan uit meerdere dia’s. Ook kan er gesproken woord en muziek worden ingevoerd. Door het gebruik van hyperlinks kan men per dia verwijzen naar andere presentaties en filmpjes.

Powerpoint kent als software een ontwikkelgeschiedenis van versies 1 – 7, editie 97, 2000, XP, 2003, 2007, 2010 en 2013. Voor opname in het e-depot worden enkel de versies vanaf 97 toegestaan.

Vanaf Microsoft Office 2007 is het standaard output-formaat van MS Powerpoint gebaseerd op onder meer het open OOXML-formaat. Daarvoor was het een gesloten (proprietair) formaat, zoals Powerpoint 2000, 97-2002 en XP, waarvan de structuur niet openbaar bekend is.

Daarnaast kunnen Powerpoint-bestanden multimedia content bevatten, zoals filmpjes, geluidsfragmenten en 3D-content. Het formaat hiervan kan verouderd raken, doordat deze ‘externe’ applicaties toekomstig niet langer bruikbaar zullen zijn.

Open alternatieven met zijn PDF en OpenDocument-formaten (.odp).

Omdat Microsoft Office Powerpoint marktleider is op het gebied van presentatie-software, alternatieven minder ver uitontwikkeld zijn en er informatieverlies op kan treden bij migratie naar andere bestanden, adviseren wij het origineel te bewaren.

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Presentatie** | **Opmerkingen** |
| **Opname** | ODP / PDF-A / PPT / PPTX | Acceptatie enkel versies vanaf circa 2000  **Op te nemen versies:**  PDF-A: Zie PDF  ODP: fmt/138, fmt/292, fmt/293  PPT: fmt/126  PPTX: fmt/215 |
| **Preservering** | geen | Preserveren indien kwetsbaar (“Just in time”) |
| **Opmerkingen** |  | Geen acties benodigd |

## Databank

Er is geen officiële open standaard voor het archiveren van databases. Er bestaan diverse alternatieven als wij ons willen beperken tot een open, valideerbaar en volledig gedocumenteerd bestandsformaat dat voldoet aan een open standaard. Een daarvan is de SIARD-suite, een freewaretool voor het omzetten van relationele databases (Oracle, Microsoft SQL-server, MySQL, DB/2 en Microsoft Access) naar het SIARD-formaat. Een andere mogelijkheid is de Database Markup Language (DBML), een ontwikkeld XML-formaat voor de opslag van de structuur en inhoud van relationele databases.

Een veel voorkomende databaseformaat is van Microsoft Access. Omdat Microsoft met Access wereldwijd marktleider is op het gebied van desktopdatabases, wordt Access veel gebruikt. Als het wordt aangeboden, verdient het de voorkeur het niet te converteren naar een andere, open standaard. Hierdoor vindt informatieverlies plaats.

Voor het e-depot lenen zich voornamelijk het bewaren van bevroren exports of ‘snapshots’ van databases. Als databases nog benaderd of geëmuleerd moeten worden, dan moet dat buiten en los van het e-depot gebeuren. Zodoende zijn er ook geen specifieke e-depottools beschikbaar voor Access. Vooralsnog laat het e-depot om bovenstaande redenen geen databases anders dan in XML formaat toe.

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Databank** | **Opmerkingen** |
| **Opname** | XML | Acceptatie XML: fmt 101 |
| **Preservering** |  | Conversie buiten e-depot naar XML m.b.v. SIARD (2) |
| **Opmerkingen** |  | Indien databases aangeboden worden dan onderzoek doen; mogelijke conversie buiten het e-depot naar XML m.b.v. SIARD |

## Vector Image

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Vector image** | **Opmerkingen** |
| **Oorspronkelijk** | SVG / EPS | SVG: fmt/91, fmt/92, fmt/413  EPS: fmt/122, fmt/123, fmt/124 |
| **Opname** | SVG / EPS |  |
| **Preservering** | geen | Preserveren indien kwetsbaar (“Just in time”) |
| **Download** | origineel |  |
| **Opmerking** |  | Overige vector bestanden in overleg, zoals bijvoorbeeld .AI. Indien een ander formaat voorkomt advies converteren naar .SVG |

## Audio

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Audio** | **Opmerkingen** |
| **Oorspronkelijk** | WAV MP3 AAC FLAC | WAV: alle versies, zie: <https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=listReport>  MP3: fmt/134  AAC: geen fmt waarde beschikbaar  FLAC: fmt/279 |
| **Opname** | WAV MP3 AAC  FLAC | Bij opname mp3 is het van belang te kijken of dit het origineel is en de kwaliteit te beoordelen op basis van gebruik. |
| **Preservering** | geen |  |
| **Download** | origineel |  |
| **Opmerkingen** |  | AAC is niet bekend in PRONOM. Dit is echter wel een standaard volgens DEN |

## Video

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Video** | **Opmerkingen** |
| **Oorspronkelijk** | MXF MPEG-4 MKV  SRT | MXF: alle versies, zie: <https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=listReport>  MPEG-4: fmt/199  MKV: fmt/569  Ondertitels (SRT) moeten embedded aangeleverd worden. (opgenomen als laag in het video bestand) |
| **Opname** | MXF MPEG-4 MKV SRT |  |
| **Preservering** | geen |  |
| **Download** | origineel |  |
| **Opmerkingen** |  | Enkel conversie naar ander formaat indien afwijkend van voorkeurs- en acceptabel formaat. Advies conversie naar MXF en hiernaast SRT indien voorzien van ondertiteling.  SRT (ondertiteling) is een apart bestand en wordt ook zo aangeleverd.  Indien ‘lossy’ wordt aangeleverd beoordelen of de kwaliteit voldoende is. |

## Website

|  |  |  |
| --- | --- | --- |
| **Formaat** | **Vector image** | **Opmerkingen** |
| **Oorspronkelijk** | WARC | fmt/1281, fmt/289 |
| **Opname** | WARC |  |
| **Preservering** | geen | Preserveren indien kwetsbaar (“Just in time”) |
| **Download** |  | De website zal middels een viewer beschikbaar gesteld worden in het e-depot |
| **Opmerking** |  | Een website kan als WARC aangeleverd worden maar ook door het e-depot ‘geharvest’ worden van het internet. |