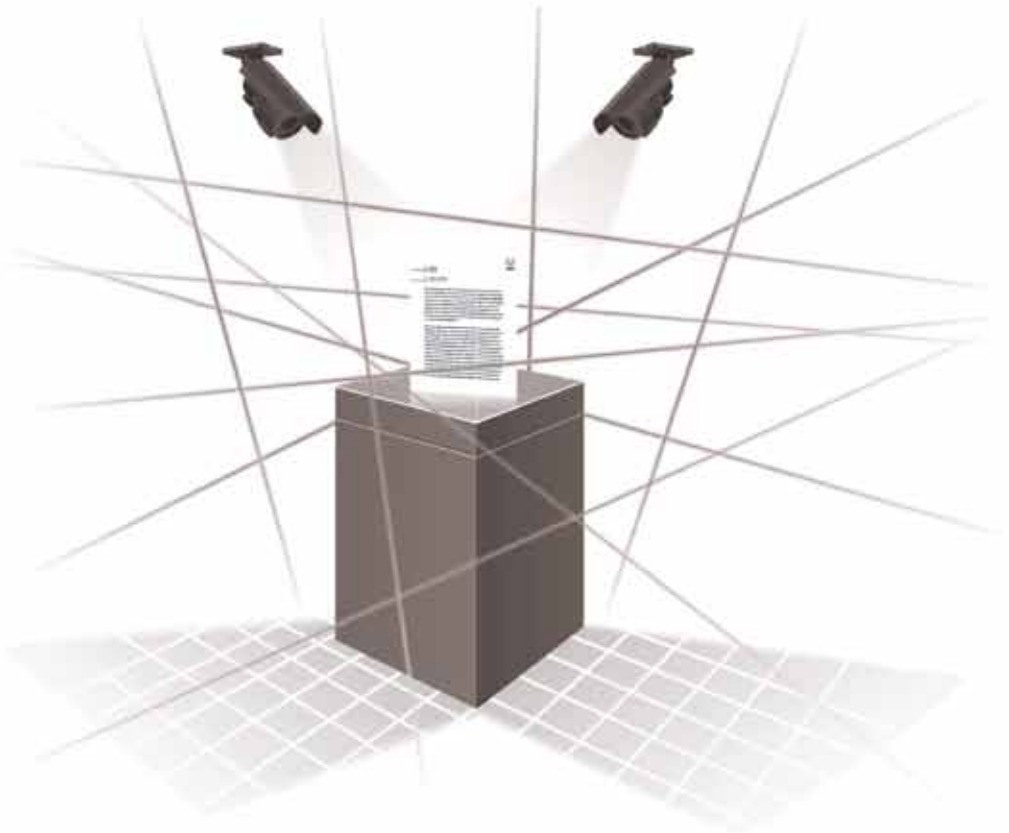


Op weg naar duurzame archivering

WHITE PAPER OVER ARCHIVEREN IN PDF



DOOR COLIN VAN OOSTERHOUT

COLOFON

Deze white paper werd gemaakt door Adobe Systems Benelux B.V.

Hoogoorddreef 54a

1101 BE Amsterdam ZO

Tel: +31 20 65 11 200

adobe_benelux@adobe.com

www.adobe.nl | www.adobe.be | www.adobe.com

© 2009 Niets uit deze uitgave mag op enigerlei wijze worden overgenomen zonder uitdrukkelijke toestemming van Adobe Systems Benelux B.V. De redactie aanvaardt geen aansprakelijkheid voor mogelijke gevolgen die zouden kunnen voortvloeien uit het gebruik van de in deze uitgave opgenomen informatie.

1	Vereisten voor elektronisch archiveren	7
1.1	Niet-reviseerbare documenten	7
1.2	Reviseerbare documenten	7
2	Het gebruik van standaardformaten voor duurzame archivering van digitale documenten	8
2.1	De drie voornaamste strategieën voor digitale archivering	8
2.2	Toekomstvast strategie	8
2.3	Wetgeving en aanknopingspunten	9
3	PDF/A	11
3.1	Wat is PDF/A?	11
3.2	De verschillen tussen PDF en PDF/A	11
3.3	Twee niveaus om aan te conformeren	12
4	Onder de motorkap van een PDF/A-document	13
4.1	Afspraken rondom PDF/A	13
5	PDF/A-software	17
5.1	PDF/A-creatie	17
6	Praktische implementatie van een PDF/A-strategie	19
7	PDF/A in relatie tot andere open documentstandaarden	21
7.1	ODF	21
7.2	TIFF	22
7.3	XML	22
8	De toekomst van PDF/A: PDF/A-2	25
	Conclusie	26



Duurzaam, integer en toegankelijk. Het zijn de primaire kwaliteitseisen voor archieven. Als bron van informatie over het handelen van mensen en organisaties zijn archieven niet alleen vandaag relevant, maar ook morgen, overmorgen en lang daarna. Gewenste informatie moet gevonden kunnen worden en als lezer moet je de inhoud kunnen vertrouwen.

Voor een moderne archivaris vormt dit de kern van het beroep. Niet meer het beheren en opruimen van oude documenten, maar het vormgeven van de informatiehuishouding binnen (overheids)organisaties. De moderne archivaris heeft een dynamische functie en opereert op het snijvlak van bedrijfsproces, ICT en archivering. De uitdagingen zijn legio. De gewenste kwaliteitseisen zijn immers in de nieuwe digitale wereld niet vanzelfsprekend, maar wel haalbaar. Door te kiezen voor een duurzaam documentformaat, waardoor documenten nu en in de toekomst goed leesbaar en uitwisselbaar zijn, zet u een stap op die weg.

In het najaar van 2008 werd ik door het Forum Standaardisatie gevraagd voorzitter te zijn van een expertgroep die zich zou buigen over PDF/A, een relatief nieuwe ISO standaard die in 2005 het licht zag. De centrale vraag vanuit het Forum was of PDF/A bij zou dragen aan de digitale samenwerking (interoperabiliteit) tussen bedrijven, burgers en overheden. PDF/A draagt niet alleen bij aan de uitwisselbaarheid op korte termijn, maar is ook toekomstgericht. PDF/A werd op de 'Lijst met Open standaarden' opgenomen.

Na de plaatsing op de Lijst met Open Standaarden heeft het gebruik van deze standaard een snelle vlucht genomen. Veel organisaties doen momenteel hun eerste ervaringen op met het implementeren van PDF/A en moeten hier tegelijkertijd beleid omheen ontwikkelen. Dit doet veelal een beroep op het vermogen om hieromheen beleid te ontwikkelen. Met natuurlijk veel vragen tot gevolg.

Om bij te dragen aan de kennis die nodig is om succesvol met PDF/A aan de slag te gaan, heeft Adobe deze white paper geschreven. In dit document wordt op een heldere manier uitleg gegeven over de thema's rondom PDF/A. Wat kunt u er wel mee, en wat niet? Wat is de achtergrond van PDF/A en hoe ziet de toekomst er uit? Met dit document als leidraad kunt u de eisen aan de 21e eeuwse informatiehuishouding met opgeheven hoofd tegemoet treden!



Erika Hokke

Adviseur organisatie en informatisering bij Digital Display, voorzitter van het PDF/A expertcommittee

INTRODUCTIE

Een vraag waar records managers voortdurend door worden geplaagd, is welk bestandsformaat zij het beste kunnen gebruiken voor het langdurig archiveren van elektronische documenten. Deze vraag wordt gecompliceerd door veel factoren, waaronder het brede spectrum aan archiveringssituaties en speciale vereisten, de voortdurende evolutie van bestaande bestandsformaten en de uitvinding van nieuwe soorten bestandsformaten. De moeilijkheden die ontstaan bij het maken van plannen voor de juiste hardware, software en media om ervoor te zorgen dat ook in de toekomst blijvend met deze documenten kan worden gewerkt, spelen hierbij een belangrijke rol.

Deze white paper gaat dieper in op mogelijkheden om PDF/A in te zetten voor het langdurig archiveren van elektronische documenten. Hierbij wordt gekeken naar de vereisten voor elektronisch archiveren, bestaande strategieën voor digitale archivering en de wetgeving die geldt voor het gebruik van standaardformaten. Vervolgens wordt de geschiktheid en het toepassen van Portable Document Format (PDF) en PDF/A voor verschillende taken, nu en in de toekomst, besproken. Tevens wordt een vergelijking gemaakt met andere belangrijke elektronische bestandsformaten. Deze white paper kan records managers helpen bij het maken van een weloverwogen keuze voor het juiste bestandsformaat voor het langdurig archiveren van elektronische documenten binnen hun organisatie.

1 VEREISTEN VOOR ELEKTRONISCH ARCHIVEREN

Bij elektronisch archiveren hebben we doorgaans te maken met verschillende doelen en diverse soorten materiaal. Aan dit onderscheid wordt in de praktijk vaak weinig aandacht geschonken. Om te kunnen bepalen welk bestandsformaat het meest geschikt is voor archivering zal antwoord moeten worden gevonden op een belangrijke eerste vraag: wat voor soort document moet worden gearchiveerd? Voor traditionele tekstdocumenten zijn twee algemene categorieën elektronische documenten te onderscheiden: niet-reviseerbare en reviseerbare documenten.

1.1 NIET-REVISEERBARE DOCUMENTEN

Wanneer het de taak is om bestaande documenten onveranderd maar in elektronische vorm te bewaren, zijn zogeheten niet-reviseerbare documentformaten noodzakelijk. PDF/A is een voorbeeld van een niet-reviseerbaar documentformaat. Hiervan verwachten we dat het een document op een elektronische manier representeert, waarbij de visuele integriteit gewaarborgd is en het document niet gemakkelijk is te veranderen. In deze categorie bevinden zich onder meer zakelijke transacties en juridische documenten. Over het algemeen is hierbij het doel de documenten te behouden voor toekomstige referentie, op een manier waarop we ook papier bewaren. De nadruk ligt hierbij dus op het behoud van de visuele integriteit. Deze documenten kunnen ontstaan zijn door ze te scannen of kunnen 'digitaal geboren' (digital born) zijn. De mogelijkheid om deze documenten aan te passen is geen vereiste.

1.2 REVISEERBARE DOCUMENTEN

Wanneer het noodzakelijk is om een soort 'brondocument' te behouden dat kan worden aangepast om er op de lange termijn afgeleide documenten van te maken, dan is een documentformaat nodig waarvan de bewerkbaarheid gewaarborgd is. Het ODF-bestandsformaat is een voorbeeld van een reviseerbaar documentformaat. Het bewaren van de exacte representatie weegt hierbij veel minder zwaar dan de bewerkbaarheid van het document.

2 HET GEBRUIK VAN STANDAARDFORMATEN VOOR DUURZAME ARCHIVERING VAN DIGITALE DOCUMENTEN

Naast de soort documenten speelt ook de te volgen strategie voor lange termijn bewaring van elektronische documenten een belangrijke rol bij het bepalen van het juiste bestandsformaat. Daarbij is het van belang dat documenten ook in de toekomst te raadplegen zijn en dat het gekozen formaat aansluit bij wetgeving voor het gebruik van standaardformaten.

2.1 DE DRIE VOORNAAMSTE STRATEGIEËN VOOR DIGITALE ARCHIVERING

Momenteel zijn er drie belangrijke strategieën die bruikbaar zijn om het vraagstuk rond de lange termijn bewaring van elektronische documenten op te lossen:

- **Migratie;** Bij migratie worden oudere bestandsformaten met nieuwe software geopend en weer weggeschreven in een up-to-date bestandsformaat. Dit is bijvoorbeeld het geval als een Microsoft Office 2000 document wordt geopend en opnieuw wordt opgeslagen met Microsoft Office 2007.
- **Emulatie;** Bij emulatie wordt niet de originele hard- en software bewaard maar wordt het vereiste platform op een toekomstige computerconfiguratie gereconstrueerd, zodat de computerbestanden in hun oorspronkelijke formaat raadpleegbaar zijn. Emulatie kan op diverse niveaus worden toegepast. Men kan computerhardware, besturings-systemen, specifieke software of een combinatie van dit alles nabootsen.
- **Gebruik van standaard documentformaten;** Bij deze strategie wordt geprobeerd zoveel mogelijk gebruik te maken van al dan niet open standaarden om de lange termijn bewaring van elektronische documenten te waarborgen.

2.2 TOEKOMSTVASTE STRATEGIE

Voor migratie geldt enerzijds dat deze strategie 'verlies' met zich mee brengt, in de zin dat het document er niet altijd meer exact hetzelfde uitziet en dat informatie verloren kan gaan. Anderzijds brengt deze strategie een relatief omvangrijke beheerslast voor de archiefinstelling met zich mee. Voor emulatie geldt dat de praktische uitvoerbaarheid ervan door vele archiefexperts in twijfel wordt getrokken. Eén van de nadelen die bijvoorbeeld aan emulatie is verbonden, is dat het technisch vrij complex is. De know-how en expertise voor ontwikkeling en onderhoud zijn binnen

archieven vaak niet aanwezig. Archieven worden daardoor afhankelijk van externe diensten en partners.

Om de raadpleegbaarheid van digitale documenten op middellange en lange termijn te maximaliseren, bestaat daarom op dit moment slechts één goede implementeerbare strategie: het gebruik van open standaarden als bestandsformaat. Dit is een toekomstvaste strategie die de beste garantie biedt op het behoud van raadpleegbaarheid.

2.3 WETGEVING EN AANKNOPINGSPUNTEN

Bij het bepalen van de strategie en daarmee de keuze voor een bestandsformaat moet rekening worden gehouden met wet- en regelgeving. Het wettelijke kader voor het gebruik van standaardformaten binnen de overheid wordt feitelijk bepaald door twee belangrijke documenten:

- Artikel 6 van de ministeriële regeling Geordende en Toegankelijke Staat Archiefbescheiden (2002). Deze regeling wordt momenteel herzien.
- Artikel 25 van het concept van de nieuwe ministeriële regeling, gebaseerd op de evaluatiecommissie (augustus 2008) die de regeling uit 2002 heeft beoordeeld.

De verwachting is dat de nieuwe Archiefregeling in de loop van 2009 van kracht wordt en dan de huidige regeling Geordende en Toegankelijke Staat Archiefbescheiden uit 2002 zal vervangen. Het advies voor deze regeling stelt dat de digitale archiefbescheiden van de overheid uiterlijk op het tijdstip van overbrenging naar een archiefbewaarplaats moeten worden opgeslagen, volgens een standaard die voldoet aan de volgende eisen:

- het opslagformaat is gedocumenteerd;
- het is een open standaard, tenzij dat onmogelijk is;
- compressie is alleen toegestaan indien aantoonbaar geen informatieverlies optreedt;
- het opslagformaat kent geen encryptie.

Er zijn diverse standaardformaten die (deels) voldoen aan de criteria die worden gesteld in de ontwerparchiefregeling. TIFF, PDF(/A), ODF en XML zijn enkele van de meer populaire formaten die hierin passen, waarbij het relatief nieuwe PDF/A aan een sterke opmars bezig is. Daarom zullen we in de volgende paragraaf verder ingaan op dit documentformaat en vervolgens verderop in deze white paper de toepasbaarheid ervan vergelijken met andere documentformaten.



3.1 WAT IS PDF/A?

PDF/A is een elektronisch bestandsformaat dat bedoeld is om documenten te presenteren, waarbij zowel de inhoud als ook de visuele verschijningsvorm en de structuur van een document over een lange periode en onafhankelijk van systemen of software te waarborgen is. Geen ander documentformaat heeft dezelfde ingebouwde mogelijkheden als PDF om zowel de inhoud, structuur en representatie te bewaren, onafhankelijk van besturingssystemen of apparaten. Om ontsluiting in de toekomst mogelijk te maken, worden in een PDF/A document alle zaken die toekomstige representatie in de weg kunnen zitten, volledig uitgesloten.

Veel organisaties gebruiken al een 'normale' PDF voor archivering. Het probleem met een 'normale' PDF is echter dat de functierijke natuur ervan problemen kan veroorzaken bij het langdurig bewaren van informatie. Een voorbeeld: PDF-documenten zijn niet noodzakelijkerwijs autonoom, aangezien sommige bestanden een afhankelijkheidsrelatie kennen met systeemfonts of met andere inhoud die van buiten het bestand opgehaald wordt. Omdat technologie voortdurend aan verandering onderhevig is, kunnen deze afhankelijkheden ervoor zorgen dat informatie verloren gaat. Omdat er daarnaast ook veel PDF-ontwikkelaars op de markt opereren, is er inconsistentie in het bestandsformaat gekomen. Dat betekent dat toekomstige migratie van PDF-bestanden wel eens moeilijk kan zijn, omdat records managers niet per definitie weten wat er 'onder de motorkap' zit. Om ervoor te zorgen dat documenten toegankelijk kunnen blijven over een langere tijdsperiode, is dan ook een lange termijn oplossing nodig.

PDF/A biedt deze oplossing. Het bestandsformaat heeft zijn oorsprong in de Verenigde Staten, via een samenwerkingsverband tussen AIIM en NPES (The Association for Suppliers of Printing, Publishing, and Converting Technologies). In 2002 startte hierdoor een gezamenlijke werkgroep (WG5) waarin, naast vertegenwoordigers van de technische committees van ISO, diverse vertegenwoordigers van bibliotheken, archivarissen, PDF-softwareontwikkelaars, vertegenwoordigers van de overheid, grafische experts en anderen samenwerkten om de PDF/A-standaard te definiëren. De standaard werd wereldwijd aangenomen in juni 2005.

3.2 DE VERSCHILLEN TUSSEN PDF EN PDF/A

De PDF/A-1 standaard is gebaseerd op Adobe's PDF Referentie 1.4. Het specificeert hoe een subset van PDF-componenten kan worden gebruikt voor het ontwikkelen van software waarmee een bepaalde 'smaak' PDF-documenten kan worden gemaakt, getoond en anderszins verwerkt, en die meer geschikt is voor langdurige archivering dan een traditioneel

PDF-document. PDF/A-1 is gericht op het langdurig behouden van de statische visuele verschijningsvorm van elektronische documenten. Daarnaast voorziet de standaard in de toekomstige behoeftes ten aanzien van toegang en migratie. Dit wordt gedaan door een raamwerk te bieden waarbij enerzijds metadata wordt ingebed en anderzijds de logische structuur en semantische eigenschappen worden gedefinieerd. PDF/A-1-bestanden zijn meer autonoom, meer zelfbeschrijvend en meer apparaatafhankelijk dan traditionele PDF 1.4-bestanden.

3.3 TWEE NIVEAUS OM AAN TE CONFORMEREN

PDF/A-1 kent twee verschillende niveaus om aan te conformeren: PDF/A-1a en PDF/A-1b. PDF/A-1a is bedoeld voor digitaal geboren documenten, waarbij de structuur van het document binnen PDF is opgenomen en PDF/A-1b is juist bedoeld voor het behoud van de verschijningsvorm van het document zonder dat de structuur daarbij van belang is.

Het niveau PDF/A-1a voldoet aan alle eisen zoals die gesteld zijn in de PDF-referentie die is aangepast in het kader van de ISO 19005-specificatie, en stelt dat alle structurele en semantische eigenschappen moeten worden bewaard. Het maakt gebruik van zogenaamde 'Tagged PDF' (ook wel bekend als gelabelde PDF). Deze labels beschrijven de natuurlijke leesvolgorde van het PDF-document. Daarnaast wordt ook Unicode gebruikt. Unicode is een internationale standaard voor de identificatie van grafische tekens en symbolen. Zowel 'Tagged PDF' als Unicode zorgen ervoor dat de logische structuur en inhoud van de tekst in de natuurlijke leesvolgorde wordt behouden.

Bij het niveau PDF/A-1b geldt dat de vereisten zo zijn opgesteld dat minimaal kan worden voldaan aan het behouden en verzekeren van de visuele integriteit van elektronische documenten.

PDF/A behelst twee elementen: een aan te passen viewer die de afspraken in een PDF/A-document respecteert en daarnaast een aantal 'slimme' afspraken binnen een documentformaat. Een aantal zaken wordt ook niet opgelost door PDF/A. Denk hierbij aan problematiek met betrekking tot de opslag van data. Voor een PDF/A-document maakt het niet uit of dit op cd-rom, DVD, harddisk of een speciaal storagestelsel wordt opgeslagen. Ook migratie van hardware valt buiten de scope van PDF/A en hetzelfde geldt voor migraties van besturingssystemen. Ten slotte doet PDF/A ook geen 'uitspraak' over documentmanagement. Dit laatste is een interessant punt, omdat zaken zoals de toegang tot het document, de omgang met metadata en de beveiliging wel op het niveau van de records managementomgeving geregeld moeten worden.

4.1 AFSPRAKEN RONDOM PDF/A

Om de representatie van een PDF/A-document in de toekomst te waarborgen, zijn afspraken opgesteld waarin bepaalde elementen in een PDF/A verboden zijn, andere zaken verplicht zijn gesteld en weer andere elementen worden aanbevolen. Een compleet overzicht van deze afspraken is te vinden in de PDF/A-specificatie die tegen nominale kosten verkrijgbaar is bij het Nederlands Normalisatie Instituut (NEN). Onder de belangrijkste afspraken vallen:

1. Kleurenruimtes;

PDF/A-documenten mogen alleen zogenaamde apparaatafhankelijke kleurenruimtes gebruiken zoals CalGray, CalRGB, of Lab. Alle kleurenruimtes moeten worden ingebed en conformeren aan de ICC (International Color Consortium) specificatie.

2. Compressie;

PDF/A-documenten bevatten momenteel idealiter alleen verliesvrije (lossless) compressie-algoritmes waarop geen intellectueel eigendom rust. Deze algoritmes zijn bedoeld om documenten kleiner in bestandomvang te maken. Het gebruik van LZW-compressie is verboden, het gebruik van ZIP-compressie wordt aanbevolen.

3. Externe verwijzingen;

Voor de doelstellingen van langdurige archivering is het verplicht dat het document volledig autonoom is, dat wil zeggen dat er geen externe afhankelijkheden mogen bestaan. Hierdoor mogen PDF-functies die te maken hebben met externe referenties niet worden gebruikt. Onder deze functies vallen acties die externe applicaties aanroepen en acties die JavaScript (eenvoudige scripting taal) gebruiken. Het gebruik van JavaScript kan een externe afhankelijkheid creëren en in de weg gaan zitten bij het betrouwbaar representeren van een document. Hyperlinks die verwijzen naar bijvoorbeeld een website of een ander document, vormen een uitzondering.

Strikt genomen zijn deze hyperlinks nog wel in het document aanwezig maar wordt het 'gedrag' (het volgen van de hyperlink) weggehaald in een PDF/A-conformerende viewer. Dat wil zeggen: de bestemming van een hyperlink in een PDF/A-document is nog wel te zien maar het klikken op de link leidt tot geen enkele activiteit.

4. Fonts;

Alle fonts die worden gebruikt in een document, inclusief standaardfonts, moeten (waar mogelijk) worden ingebed, zodat het document op een later moment in de tijd altijd betrouwbaar kan worden getoond en er uitziet zoals de oorspronkelijke auteur het ooit bedoeld heeft. Om de bestandsomvang te verkleinen, kunnen de ingebedde fonts ook als subset worden opgenomen.

5. Formulieren;

Om het consistent tonen van formulervelden te verzekeren, is iedere actie die mogelijk de visuele integriteit kan aantasten verboden. Denk hierbij aan formulervelden waar JavaScript in zit om bijvoorbeeld een datum te tonen.

6. Afbeeldingen;

Alternatieve afbeeldingen die op verschillende manieren kunnen worden gerepresenteerd, mogen niet gebruikt worden. Het gebruik van transparantie in een afbeelding is eveneens verboden, in plaats daarvan wordt aanbevolen om afbeeldingen met lagen af te vlakken (samen te voegen). Het downsamplen (terugbrengen van de resolutie) van afbeeldingen gedurende het PDF-creatieproces moet worden vermeden, omdat het kan resulteren in kwaliteitsverlies (dit is overigens geen formele PDF/A-vereiste).

7. Metadata;

Om uniform de omschrijvende, administratieve en technische metagegevens te beschrijven, moet het PDF-bestand een zogenaamde metadatastream bevatten die voldoet aan de XMP-specificatie.

8. Multimedia;

Het inbedden van multimedia-inhoud, waaronder het gebruik van geluidsannotaties, acties die gekoppeld zijn aan geluiden, filmannotaties en acties die aan filmannotaties zijn verbonden, is verboden.

9. Beveiliging;

Het document mag niet zijn voorzien van encryptie of wachtwoordbeveiliging. Deze verhinderen de toegang tot het document, waardoor ook migratie in de toekomst voor problemen zou kunnen zorgen.

10. Gebruik van digitale handtekeningen;

Het gebruik van digitale handtekeningen in archiefdocumenten vormt een uitdagend vraagstuk. Afhankelijk van de exacte archiveringsvereisten moeten documenten soms twintig jaar of langer bewaard blijven. Handtekeningen in PDF worden gebruikt om de authenticiteit en de integriteit van het document te waarborgen. De hiervoor benodigde handtekeningcertificaten zullen echter op termijn vervallen of de technologie waarmee de handtekening is gemaakt zal evolueren. Hoe kunnen de eisen rondom langdurige archivering dan toch worden gerijmd met het gebruik van digitale handtekeningen? Er zijn verschillende methodes voorhanden, maar wellicht de meest krachtige optie is internalisering van het authenticatieproces. Hierbij worden de checks ten behoeve van de geldigheid van het certificaat tesamen met het gearchiveerde document vastgelegd. Op termijn biedt deze verificatiemetadata waarschijnlijk de enige methode voor het verifiëren van de authenticiteit van het origineel. Zolang het archief zelf intact blijft, blijft de verificatiemetadata aanwezig en kan het document worden gevalideerd.



Er valt heel veel te ontdekken over zaken die onder de motorkap plaatsvinden bij het creëren van PDF/A-documenten. Toch zal een PDF/A-document uiteindelijk via specifieke instellingen in software of hardware, zoals scanners of multifunctionals, worden gemaakt of gecontroleerd. U hoeft daarom geen PDF/A-expert te zijn om er toch mee te kunnen werken. PDF/A-software zorgt hierbij voor het valideren, tonen en vooral ook het creëren van PDF/A-documenten.

5.1 PDF/A-CREATIE

Er zijn twee hoofdcategorieën software voor PDF/A-creatie: desktopsoftware en serversoftware. Daarnaast is er een aparte categorie voor hardwareapparaten. Hier gaat het om de leveranciers van scanners, kopieermachines en multifunctionals die ingebouwde mogelijkheden voor PDF/A in zich dragen. Ook leveranciers van document management systemen of postregistratiesystemen bieden tegenwoordig mogelijkheden om documenten om te zetten naar PDF/A. Het meest voorkomend zijn momenteel echter de desktop- en servertools.

Desktoptools

Deze tools bevinden zich op de desktop van een eindgebruiker. Er zijn vele gratis en niet-gratis tools om PDF's mee te creëren. De kwaliteitsverschillen zijn echter groot. Bij het archiveren van PDF/A bestanden is het gebruik van een goede tool daarom belangrijk. Voor het maken van PDF/A-bestanden worden over het algemeen drie methodes aangeboden:

- Via de lokale printerfunctie
- Via de bewaar- of exportfunctie van MS Word (of een ander tekstverwerkingsprogramma)
- Via de scanfunctie

Voor de eerste en derde methode geldt dat er altijd PDF/A-1b-bestanden worden geproduceerd. Via de tweede methode is ook de creatie van PDF/A-1a-bestanden mogelijk. Sommige desktoptools zijn geschikt om bestanden tegelijk (in grote hoeveelheden) naar PDF/A te converteren of te valideren. Toch zullen hier altijd beperkingen optreden, omdat desktoptools niet bedoeld zijn voor geautomatiseerde en grootschalige processen. In zo'n geval is het vaak beter om servertools te gebruiken.

Servertools

Serverside tools draaien vanaf een server en kunnen over het algemeen grote aantallen documenten op hoge snelheid converteren. Deze softwaretools variëren van software die allerlei input aankan en meervoudige output genereert (waaronder PDF/A), tot tools die exclusief PDF(/A)-bestanden creëren. Serverside tools maken, door hun werking, bijna altijd PDF/A-1b-bestanden.

Nadat een bestand is omgezet naar PDF/A, kan de behoefte bestaan om op ad hoc basis (via desktoptools) of op meer gestructureerde basis (via servertools) te controleren of de conversie goed is gelukt. Bij een gedeeltelijk gelukte conversie is het soms nodig om een analyse uit te voeren om vast te stellen waar er zich problemen voordoen. Tools die hier voor zorgen, kennen we onder de naam 'validators'. Validators bieden naast analysemogelijkheden soms (maar niet altijd) de mogelijkheid om eventuele problemen in een PDF/A bestand te repareren.

De vereisten van PDF/A-1 benadrukken een betrouwbare en voorspelbare manier voor het visueel integer tonen van statische documenten. Voor het bekijken en printen van een PDF/A-1-document is een viewer nodig die in staat is om de visuele representatie van een document te waarborgen. Daarnaast moet een dergelijke viewer een aantal andere zaken goed afhandelen. PDF/A-1 staat het gebruik van bepaalde interactieve elementen zoals annotaties (opmerkingen) en hyperlinks toe, maar zal van een PDF/A-conformerende viewer verwachten dat deze als inactief worden behandeld.

6 PRAKTISCHE IMPLEMENTATIE VAN EEN PDF/A-STRATEGIE

Nu duidelijk is wat de wettelijke kaders, inhoudelijke vereisten en softwaremogelijkheden voor PDF/A zijn, kunnen we verder ingaan op de mogelijkheden voor de implementatie van een PDF/A-strategie. Hierbij spelen adviezen vanuit de overheid een belangrijke rol.

Het kabinet streeft naar verbetering van overheidsdienstverlening en administratieve lastenverlichting. In 2006 is daarom het College en Forum Standaardisatie ingesteld om de digitale samenwerking tussen bedrijven, burgers en overheden te bevorderen. Het Forum Standaardisatie heeft, na een consultatieronde van een expertcommissie, PDF/A-1 op de lijst van standaarden opgenomen. In het expertadvies zijn een aantal praktische aanknopingspunten te vinden voor het implementeren van een PDF/A-strategie:

- Bij de aanschaf van ICT-diensten of ICT-producten dient een overheidsinstelling als regel te kiezen voor open standaarden of, in die gevallen dat er goede gronden zijn om dat toch niet te doen, vast te leggen welke die goede gronden zijn. Hiermee wordt voor het Rijk invulling gegeven aan het zogenoemde 'comply or explain'-principe bij het gebruik van open standaarden. Dit is in lijn met wat is aangekondigd in het Actieplan Nederland Open in Verbinding.
- Volgens het Forum Standaardisatie voldoet de standaard PDF/A-1 aan criteria met betrekking tot openheid, bruikbaarheid en potentieel. Verder stellen zij dat 'de invoering van de standaard PDF/A-1 vooral impact heeft op het creatieproces van documenten. Organisaties zullen daarom regels moeten opstellen voor de creatie van documenten, zodat deze documenten ook opgeslagen kunnen worden als PDF/A-1a-formaat. Een tweede effect is dat de standaardisatie in PDF/A-1 zorgt voor betere langdurige opslag en beschikbaarheid van documenten en vereenvoudiging van conversie- en migratietrajecten. Tot slot versterkt de invoering van PDF/A-1 de implementatie van de NEN-ISO 15489:1 (de kwaliteitsnorm van informatie- en archiefmanagement en van het Besluit Kwaliteit rijksoverheidswebsites).
- Het toepassingsgebied van PDF/A-1 zoals beschreven door het Forum Standaardisatie betreft de eindversies van documenten die organisaties creëren of ontvangen bij de uitvoering van overheidstaken. Binnen dit toepassingsgebied is een uitsplitsing naar bestanden waarbij slechts de visuele weergave van het document gerealiseerd wordt (conform PDF/A-1b), en bestanden waarbij ook structuur en semantiek correct worden gerepresenteerd (conform PDF/A-1a). Concreet betekent dit dat alle nieuwe 'digital born' documenten die worden gecreëerd door de overheid bij de uitvoering van overheidstaken volgens PDF/A-1a worden opgeslagen.

Opslag vindt volgens PDF/A-1b plaats wanneer het gaat om:

- Analoge (papieren) documenten en afbeeldingen die worden gedigitaliseerd (gescand).
- Bestaande / oude (legacy) digitale documenten die niet correct over te zetten zijn naar PDF/A-1a, omdat ze gemaakt zijn met oude software, omdat er legacy fonts in zijn opgenomen, omdat er bijvoorbeeld wiskundige vergelijkingen in staan, of om andere redenen waardoor omzetten naar PDF/A-1a niet mogelijk is.

■ PDF/A heeft dus vooral impact op het creatieproces van documenten. De toepassing van PDF/A-1a vraagt namelijk van de auteur van de documenten om gedurende het creatieproces 'nette' documenten te maken. Dit zijn documenten met een goede structuur en met correct toegevoegde metadata. Hier zit zowel een valkuil als een kans; veel organisaties zijn nog steeds niet in staat om gestructureerde documenten te maken. Soms komt dit door het ontbreken van een goede huisstijl en sjablonen, maar soms zijn er ook andere oorzaken. Heel vaak komt de situatie voor dat documenten 'digital born' zijn (en dus een uitstekende kandidaat zijn voor omzetting naar PDF/A-1a) maar dat het document vervolgens wordt geprint om er een handtekening op te plaatsen. Dit document wordt dan vervolgens opnieuw ingescand en kan dan logischerwijs alleen nog maar PDF/A-1b worden. Scannen van informatie is dus niet altijd de beste optie, al kan het door het ontbreken van het origineel of vanwege andere redenen soms onontkoombaar zijn.

■ Een goede implementatie van een PDF/A strategie biedt de kans om te zorgen voor verdere regulering van het creatieproces van documenten. Dit versterkt de implementatie van de NEN-ISO 15489:1. Een ander positief effect is dat door standaardisatie van PDF/A-1 duurzame opslag met garantie van bruikbaarheid en wereldwijde ondersteuning kan worden gegarandeerd.

7 PDF/A IN RELATIE TOT ANDERE OPEN DOCUMENTSTANDAARDEN

PDF/A is dus in veel gevallen een geschikt bestandsformaat voor het langdurig opslaan van elektronische documenten. Desalniettemin zijn er diverse andere open documentstandaarden die in bepaalde situaties ook prima gebruikt kunnen worden. De belangrijkste van deze standaarden komen in de volgende drie paragrafen aan bod en worden hierin vergeleken met PDF/A.

7.1 ODF

De afgelopen jaren zijn er nieuwe open standaarden op de markt gekomen voor het uitwisselen van bewerkbare documenten. De belangrijkste daarvan is Open Document Format. ODF (ISO/IEC 26400) is een bestandsformaat voor het maken van elektronische kantoordocumenten, zoals tekstbestanden, spreadsheets en presentaties, waarbij de documenten volledig bewerkbaar blijven. De indeling is van oorsprong gebaseerd op een XML-indeling. ODF is tussen 2002 en 2005 ontwikkeld door de Organization for the Advancement of Structured Information Standards (OASIS) en is in november 2006 een ISO-standaard geworden.

Bij het implementeren van een open documentinfrastructuur is het van belang dat overheden zich niet alleen op de technologie concentreren, maar ook op de toepassing van die technologie. Met andere woorden: het uiteindelijke doel moet niet zijn om over te stappen op zo weinig mogelijk open standaarden, maar om de juiste open standaard voor de juiste toepassing te vinden. In theorie kunnen zowel ODF als PDF worden gebruikt voor praktisch elke documentgerelateerde taak, of het nu gaat om het bewerken van niet-definitieve documenten of het uitwisselen van definitieve documenten.

Je zou kunnen beargumenteren dat één van deze standaarden voldoende is voor een open documentarchitectuur. Bij het gebruik van een bestandsformaat voor langdurige archivering is het echter met name PDF/A dat de beste kaarten in handen heeft. Het verschil met PDF zit met name in het adoptie (adoption) criterium: het gebruik van het jonge ODF- bestandsformaat is zowel wereldwijd als in de culturele erfgoedsector nog (relatief) gering. Een ander en veel belangrijker probleem van het ODF-formaat is dat een eenduidige weergave van de documentopmaak tussen verschillende viewers en omgevingen niet is te garanderen. Zo kunnen verschillen optreden in regeleindes, pagina-eindes, de positionering van elementen als afbeeldingen en tabellen, en fonts. Doordat fonts niet ingebed kunnen worden in het ODF-formaat, is een eenduidige weergave op computers waar het betreffende font niet aanwezig is onmogelijk.

7.2 TIFF

Omdat de meeste activiteiten op het vlak van documentarchivering in het verleden gericht waren op papieren documenten, is het logisch dat we deze willen converteren naar elektronische representaties voor elektronische archivering. Het scannen van papier naar computerafbeeldingen is hierbij vaak het startpunt. Het TIFF-bestandsformaat wordt hiervoor traditioneel veel gebruikt. Het probleem met TIFF is dat dit formaat een hele reeks keuzes biedt voor het representeren van afbeeldingen, waaronder RGB- en CMYK-kleurenrepresentaties en JPG, LZW, FAX en andere compressietechnologieën.

TIFF, waarvan de laatste officiële publicatie (TIFF versie 6) uit 1992 stamt, wordt vaak gebruikt zonder compressie, gebaseerd op het idee dat gegevens makkelijker kunnen worden teruggevonden en niet per ongeluk worden gecorrumpeerd (zoals het geval zou kunnen zijn als er gebruik zou worden gemaakt van verliesvrije compressie). De prijs die hiervoor wordt betaald is een grotere bestandsomvang, die ook meer archiveringsruimte zal eisen. Er is ook enig risico bij het gebruik van TIFF voor archivering, omdat het een bedrijfseigen standaard is en geen open standaard. Bovendien is het bestandsformaat buitengewoon gefragmenteerd geraakt, omdat diverse partijen de TIFF-specificatie zelf hebben doorontwikkeld.

Het scannen van papier wordt gewaardeerd omdat het het dichtst bij het bewaren van papier komt – maar dan in elektronische vorm. Hierbij ontbreken echter wel de meer uitgebreide mogelijkheden van elektronische documenten. Records managers die een voorkeur hebben voor open standaarden zouden hun gescande documenten zoveel mogelijk moeten opslaan als PDF/A-1b, waarbij in verschillende compressiemethodes kan worden voorzien en waarbij ook de doorzoekbaarheid kan worden geregeld.

7.3 XML

XML is een markeringstaal (Markup Language) voor algemene documenten waarin labels (tags) voor paragrafen, allerlei soorten lijsten, hoofdstukken en allerlei andere textuele componenten zijn bedacht. De XML-notatie wordt echter gebruikt om duizenden documentsoorten te markeren, variërend van visitekaartjes, offertes, gezondheidsdossiers en vectortekeningen tot database-ingangen en programmeertalen. Gegeven de diversiteit van speciale markeringstalen, is het zinvoller om specifiek te praten over XML voor visitekaartjes, XML voor algemene documenten of XML voor hypertext (XHTML) in plaats van de term XML in algemene zin te gebruiken.

In specifieke markten en voor specifieke toepassingen geldt zeker dat XML, via goed gedefinieerde schema's, een uitstekende manier biedt om documenten op te slaan, waarbij (hopelijk) de visuele aspecten en de inhoud volledig bewaard blijven. Het is echter de vraag of er op lange termijn viewers zijn die de XML nog kunnen representeren. Wanneer de focus van het te archiveren document ligt op de betrouwbare visuele representatie in de toekomst, verdient het gebruik van PDF/A de aanbeveling. Als de focus meer ligt op hergebruik, dan is XML wellicht een prima alternatief.

Het gebruik van specifieke XML (XMP) voor de beschrijving van metadata in een PDF-document is overigens wel sterk aanbevolen. Het gestructureerde karakter en de uitbreidbaarheid van XML zorgen ervoor dat metadatering heel nauwkeurig en toegespitst op de wensen van een organisatie kan plaatsvinden.



Hoewel PDF/A op dit moment in veel gevallen al een prima alternatief vormt voor andere open documentstandaarden zijn de ontwikkelingen rondom PDF/A deel 2 (PDF/A-2) en deel 3 (PDF/A-3) reeds in volle gang. PDF/A-2 is hierbij gebaseerd op PDF 1.6. Op basis van verzoeken van implementerende partijen en gebruikers wordt overwogen in de toekomst de volgende onderdelen aan PDF/A toe te voegen:

- JPEG 2000 beeldcompressie
- Meer geavanceerde ondersteuning voor digitale handtekeningen
- OpenType fonts
- 3D-informatie
- Audio/video-inhoud
- Consistentie met PDF/X (standaard voor drukwerk), PDF/E (standaard voor CAD-documenten) en PDF/UA (standaard voor toegankelijkheid voor visueel gehandicapten)

Het zit in de planning dat toekomstige onderdelen van PDF/A zullen worden geschreven op zo'n manier dat oudere onderdelen van PDF/A zullen aansluiten op nieuwere onderdelen van deze standaard. Op deze manier kan PDF/A de ondersteuning blijven bieden om het document tot in lengte van dagen te kunnen archiveren. Beslissingen over wat er daadwerkelijk in de nieuwe onderdelen zal worden opgenomen, liggen uiteindelijk bij de ISO.

CONCLUSIE

Er zijn diverse standaardformaten waarmee elektronische documenten langdurig gearchiveerd kunnen worden. De verschillende formaten hebben hierbij allemaal hun sterke en zwakkere punten. Voor duurzame archivering is het van belang dat een opslagstandaard wordt gebruikt die aansluit bij het soort materiaal en het doel van de archivering. Daarnaast moet het bestandsformaat passen in de gekozen archiveringsstrategie en voldoen aan de geldende wetgeving. Alleen dan zijn digitale documenten ook in de toekomst op de juiste manier toegankelijk.

PDF/A-1 zal een belangrijke bijdrage leveren aan de langdurige opslag en toegankelijkheid van digitale documenten. De standaard voldoet aan criteria voor openheid, bruikbaarheid en potentieel zoals het Forum Standaardisatie die heeft gesteld en biedt organisaties de kans om te zorgen voor verdere regulering van het creatieproces van documenten. Daarbij is het bestandsformaat volop in ontwikkeling. Op basis van feedback van de gebruikers zullen in de toekomst nieuwe onderdelen aan PDF/A worden toegevoegd. Op deze manier kan PDF/A ook aan nieuwe eisen voor het archiveren van digitale documenten blijven voldoen.

Desalniettemin is het belangrijk om te onthouden dat een documentformaat alleen nooit een archief maakt. Er zijn nog zoveel andere overwegingen betrokken bij het goed definiëren van een archiveringsproces. Het document- (bestands-)formaat is slechts een enkel onderdeel van een goed ingerichte archiveringsstrategie, zoals bijvoorbeeld ISO 15489-deel 1.

Als het goed is, heeft u de whitepaper inmiddels gelezen en heeft u er hopelijk veel interessante nieuwe kennis uit opgedaan. Dit boekje heeft een zekere attentiewaarde waardoor u het ongetwijfeld nog een tijd bij u zult houden. Tot het moment waarop dit papieren document onder een andere stapel documenten verdwijnt. Daarna is het nog een kwestie van tijd totdat u besluit om dit document, eventueel tezamen met andere documenten, weg te gooien. Het is uw persoonlijke selectiecriteria dat u zal doen besluiten zich van dit document te ontdoen. Mogelijk wordt uw besluit om het papieren document weg te gooien, vergemakkelijkt door het feit dat er ook een elektronische versie van dit document bestaat.

Voor het bewaren van elektronische documenten, bestaan eveneens selectiecriteria. Wanneer de keuze gemaakt wordt om een document langdurig te bewaren, dan is niet ieder documentformaat even geschikt om dat in te doen. PDF/A-1 is juist ontwikkeld om u die mogelijkheid wel te geven. Nadat PDF/A-1 eerder al in 2007 was verankerd in de Nederlandse webrichtlijnen voor open standaarden en toegankelijkheid, werd PDF/A-1 in het najaar van 2008 via het Forum Standaardisatie op de standaardenlijst van de Nederlandse overheid geplaatst. Veel overheidsorganisaties beginnen inmiddels hun eerste ervaringen met PDF/A-1 op te doen, soms met vallen en opstaan.

Niet ieder type document dat in de uitvoering van overheidstaken tot stand komt, is perse geschikt voor omzetting naar PDF/A-1. In het geval van een CAD-document, is het in de conversie naar PDF, bijvoorbeeld de bedoeling om "lagen", maatvoering, etc. Te behouden, iets wat in PDF/A-1 niet mogelijk is. Door de toenemende "mash-up" van documenten, waarbij ook steeds meer multimedia aan documenten wordt toegevoegd, is PDF/A-1 in voorkomende gevallen ook niet altijd het meest geschikte formaat. Dat is de reden waarom momenteel een procedure in gang is gezet om ook PDF 1.7 (ISO 32000) naar de standaardenlijst te brengen. Uiteindelijk zal de interoperabiliteit van gegevens hierdoor alleen maar groter worden en kunnen ook specifieke documentformaten worden gebruikt om specifieke doelen te ondersteunen: PDF/A-1 voor langdurige archivering en PDF 1.7 voor de uitwisseling van de "functierijkere" documenttypes.

Wat er in de toekomst nog verder zal veranderen op het gebied van documentstandaarden is lastig te voorspellen. De trends tekenen zich in ieder geval wel al in grote lijnen af. Adobe zal er naar streven om samen met partners nu en in de toekomst een actieve bijdrage op basis van seminars, whitepapers e.d. te blijven leveren aan de vraagstukken rondom elektronische documenten.



Colin van Oosterhout

Business Development Manager Adobe Systems Benelux
(cvanoost@adobe.com)

