XML-bestandsformaten voor Office-applicaties: duurzame belofte ?

Geert-Jan van Bussel

Samenvatting

Sinds het bestaan van XML wordt het gezien als een middel voor interoperabiliteit van documenten over platformen heen en als een middel voor het duurzame bestaan van documenten, dankzij het feit dat het een 'open' standaard is. Dat argument is vooral gangbaar in overheidsorganen waar de duurzame bewaring in de tijd over platformen heen een belangrijk item is, gezien de verantwoordelijkheid van deze organen archieven te vormen voor toekomstig cultuur-historisch onderzoek. Het gebruik van XML zou documenten op lange termijn toegankelijk te houden, mits de afgeleide XML-schema's voor de gegenereerde bestandsformaten de algemene karaktertrekken van XML blijven behouden. In dit artikel analyseren we de gangbare XML-bestandsformaten voor Office-applicaties juist op die argumenten: zijn ze open, interoperabel en duurzaam.

I. Inleiding

Een van de meest in het oog lopende informatietechnologische kwesties van vandaag de dag is de strijd die is losgebarsten tussen industriële grootmachten als IBM, Sun, Microsoft en Google (met in het voetspoor daarvan een hele reeks kleinere bedrijven en organisaties) over het bestandsformaat waarin vooral ongestructureerde informatie in Office-applicaties zou moeten worden opgeslagen. De term ongestructureerde informatie verwijst naar de enorme massa (vooral digitale) informatie die geen vaste gegevensstructuur heeft (of een structuur die niet makkelijk leesbaar is door een computer). Voorbeelden zijn audio- en videobestanden. Ook ongestructureerde tekstbestanden als email of tekstverwerkingsdocumenten worden tot de ongestructureerde informatie gerekend. Zelfs hoog gestructureerde teksten kunnen daaronder vallen: als de structuur met name gericht is op de layout van een document en niet op het uitvoeren van meer complexe taken met de content van het document. In 2003 schatte Merrill Lynch dat meer dan 85 % van alle bedrijfsinformatie ongestructureerd was[1]. We gaan hier verder niet in op de vraag of het onderscheid tussen ongestructureerd en gestructureerd niet meer te maken heeft met het feit dat ongestructureerde informatie niet in een gestructureerde database opgenomen is[2].

II. XML

XML is voortgekomen uit de Structured Generalized Markup Language (SGML). Dit is een metataal waarin markup-talen kunnen worden gedefinieerd, die het mogelijk maken om ongestructureerde informatie, met name documenten, te structureren en te formatteren. Dat gebeurt vanuit het principe dat representatie van de structuur en de content zelf gescheiden moeten worden van de eisen van informatieverwerking. SGML is primair ontwikkeld om het mogelijk te maken digitaal aangemaakte documenten te structureren en te delen in grote projecten binnen de overheid, de juridische dienstverlening en de vliegtuigindustrie. De meeste van deze documenten dienen vaak tientallen jaren bewaard te blijven. Het feit dat SGML gegevensuitwisseling faciliteert over verschillende systemen en platformen maakt het daarvoor tot een geschikt middel. De complexiteit van SGML heeft een breed gebruik altijd in de weg gestaan[3].
XML (eXtensible Markup Language)[4] is een dialect van SGML, dat is onwikkeld om het mogelijk te maken om SGML te gebruiken, verwerken en ontvangen op het World Wide Web. De oorsprong van de taal is te dateren tot 1996, als gevolg van de frustratie met de toepassing van SGML op het Internet. Rondom SGML waren een aantal standaarden ontstaan die uiterst moeilijk te implementeren

[1] R. Blumberg, S. Atre, 'The problem with unstructured data', *DM Review Magazine*, februari 2003, http://www.dmreview.com/article_sub.cfm?articleId=6287 (geraadpleegd op 10 november 2007).

[2] J. Berkus, 'Database soup', 2005, http://blogs.ittoolbox.com/database/soup/archives/unstructured-data-as-an-oxymoron-5588 (geraadpleegd op 10 november 2007).

[3] Voor een uitgebreide introductie in SGML zie: http://etext.lib.virginia.edu/standards/tei/teip4/index.html (geraadpleegd op 10 november 2007).

[4] Zie voor de achtergronden van XML: http://www.itwriting.com/xmlintro.php (geraadpleegd op 10 november 2007). Ook: http://etext.lib.virginia.edu/standards/tei/teip4/index.html http://xml.coverpages.org/xml.html (geraadpleegd op 10 november 2007).

waren en die weinig belangstelling wekten buiten de specialistische SGML-kringen. We spreken dan over standaarden als SGML zelf, DSSSL[5] (het transformatie framework voor presentatie van documenten) en HyTime[6] (het framework voor hypertext-presentatie). XML vereenvoudigde de implementatievereisten, met de specifieke bedoeling om de toepassing van markuptalen op Internet toe te passen.
XML heeft als voornaamste doel het faciliteren van de uitwisseling van gegevens over verschillende systemen en platformen. Het voordeel van XML is dat de taal zowel voor een mens als voor een machine leesbaar is. XML is een gedefinieerde wijze van gestructureerde vastlegging van gegevens. In XML gaat het dus vooral om de structuur van de informatie. XML wordt veel gebruikt om gegevens via het Internet te verzenden.
Door het toevoegen van semantische constraints kunnen applicatietalen in XML worden geimplementeerd. Deze betreffen bijvoorbeeld XHTML[7], MathML[8], GraphML[9], Scalable Vector Graphics[10], RSS[11], MusicXML[12] en nog duizenden andere.
Daarnaast wordt XML soms gebruikt als een specificatietaal voor dergelijke applicatietalen. Door de namen, toegestane hiërarchie en betekenis van de elementen en attributen open te laten en definieerbaar door een specifiek ontwikkeld model (een XML-schema of DTD (Document Type Definition)), biedt XML een syntactische basis voor het ontwikkelen van dergelijke applicatietalen.

[5] Voor DSSSL: http://xml.coverpages.org/dsssl.html (geraadpleegd op 10 november 2007).
[6] Voor HyTime: http://xml.coverpages.org/hytime.html (geraadpleegd op 10 november 2007).
[7] Voor XHTML 1.0: http://www.w3.org/TR/xhtml1/. Voor XHTML 1.1.: http://www.w3.org/TR/2001/REC-xhtml11-20010531/. Voor XHTML 2.0 (concept): http://www.w3.org/TR/xhtml2/ (alle geraadpleegd op 10 november 2007)
[8] Voor MathML: http://www.w3.org/Math/ (geraadpleegd op 10 november 2007).
[9] Voor GraphML: http://graphml.graphdrawing.org/specification/ (geraadpleegd op 10 november 2007).
[10] Voor SVG: http://www.w3.org/Graphics/SVG/ (geraadpleegd op 10 november 2007).
[11] Voor RSS: http://www.rssboard.org/rss-specification (geraadpleegd op 10 november 2007).
[12] Voor MusicXML: http://www.recordare.com/xml.html (geraadpleegd op 10 november 2007).

Een XML-schema of DTD is een beschrijving van een type XML-document, voornamelijk uitgedrukt in termen van constraints voor de structuur en de content van documenten van dat type, nog extra gedefinieerd boven de basisbeperkingen die door XML zelf worden afgedwongen. De syntax van dergelijke applicatietalen is dwingend: documenten moeten voldoen aan de algemene regels van XML, waardoor wordt verzekerd dat alle XML-gebruikende software minimaal in staat is om de geboden informatie te lezen en de relatieve structuur van de informatie te begrijpen[13].
XML-gegevens kunnen naar andere bestandsformaten worden geconverteerd, zoals HTML of PDF. Ze kunnen echter ook naar een ander XML-document met een andere structuur worden overgebracht. De XML-tags zijn in principe vrij te kiezen; de applicatietalen vormen dan ook een gemeenschappelijke standaard voor de uitwisseling van gegevens.
XML is een open standaard, die door het World Wide Web Consortium wordt aanbevolen. Dit Consortium heeft allerlei 'standaarden' aangaande XML ontwikkeld en/of uitgegeven.

III. Voor- en nadelen van XML.

Een documentair bestandsformaat is een tekst of een binair bestand[14] voor de opslag van documenten door computers. Er bestaat een enorme hoeveelheid van deze bestandsformaten, die veelal niet met elkaar kunnen samenwerken (incompatibel zijn). Er is een consensus dat XML de basis dient te zijn voor toekomstige documentaire bestandsformaten. Die consensus is met name geba-

[13] Voor het XML-schema zie: http://www.w3.org/TR/xmlschema-1/ en http://www.w3.org/TR/xmlschema-2/. Voor het DTD zie: http://www.w3schools.com/dtd/default.asp (alle geraadpleegd op 10 november 2007).
[14] Een binair bestandsformaat is gebaseerd op een binair numeriek systeem, waarbij numerieke waarden worden uitgedrukt door middel van twee symbolen, veelal 0 en 1. Het is een computerbestand dat elk type gegevens kan bevatten, binair gecodeerd voor opslag en verwerking. Binaire bestanden die alleen tekstuele gegevens bevatten, zonder enige layout-data, worden 'plain text files' genoemd. Als een zekere contradictie worden deze 'plain text files' over het algemeen niet als binaire bestanden gezien, juist omdat ze alleen maar textuele gegevens bevatten.

seerd op een aantal voordelen die XML biedt, met name:

1. het is een formaat dat zowel door mensen als door computers kan worden gelezen, al blijft het voor mensen niet eenvoudig[15];
2. het is zelf-documenterend;
3. het ondersteunt Unicode en kan daardoor informatie in ongeacht welke taal communiceren;
4. het is gebaseerd op internationale standaarden;
5. het heeft een hiërarchische structuur die geschikt is voor (bijna) alle soorten documenten;
6. het is betrekkelijk eenvoudig om metadata toe te voegen aan de in XML gegenereerde documenten;
7. het is platform-onafhankelijk en derhalve relatief immuun voor veranderingen in technologie;
8. het is relatief eenvoudig om voorwaartse en achterwaartse compatibiliteit te handhaven ondanks veranderingen in DTD of XML Schema.

Het zijn vooral deze laatste kenmerken die het formaat associëren met digitale duurzaamheid, vooral in overheidskringen.

De nadelen die aan XML kleven worden over het algemeen onderbelicht. Die nadelen kunnen echter consequenties hebben voor de informatievoorziening. De belangrijkste nadelen van het gebruik van XML zijn:

1. de XML-syntax is overvloedig en groot vergeleken met de binaire afbeelding van dezelfde gegevens[16], wat de efficiency van applicaties beïnvloedt door meer opslag, grotere bandbreedte voor verspreiding en hogere verwerkingskosten[17];
2. er bestaat nauwelijks ondersteuning voor data types als 'integer', 'string', 'boolean', 'data' enz., alhoewel ontwikkelde XML Schemata deze ondersteuning wel kunnen inbouwen[18];
3. de hiërarchische modellering van de representatie is veel beperkter dan de relationele of object georiënteerde modellering[19];
4. het zelf-documenterende karakter van XML is beperkt, waardoor de interoperabiliteit tussen verschillende systemen minder groot is dan verwacht[20];
5. de authenticiteit van de content van een document wordt door het muteerbare karakter van XML niet gewaarborgd.

IV. De bediscussieerde XML-bestandsformaten.

Er zijn vele XML-bestandsformaten[21], maar slechts enkele voor Office-applicaties: OpenDocument Format (ODF), Office Open XML (OOXML) en Uniform Office Format (UOF). We richten ons hier op de eerste twee. UOF staat in de discussie aan de zijlijn, maar zal wellicht, gezien de Chinese oorsprong, in de komende jaren een belangrijker plek gaan innemen. UOF beschikt over een conversiemogelijkheid naar ODF (en omgekeerd). Er wordt gewerkt aan een conversiemogelijkheid voor OOXML.

OpenDocument Format (ODF)

ODF is door de ISO/IEC gestandaardiseerd als ISO/IEC 26300, Open Document Format for Office Applications[22]. Het is een bestandsformaat voor electronische Office-documenten, zoals spreadsheets, grafieken, presentaties en tekstverwerkingsdocumenten. De standaard is ontwikkeld door een Technisch Committee[23] van de Organiza-

[15] World Wide Web Consortium, 'XML in 10 points', http://www.w3.org/XML/1999/XML-in-10-points (geraadpleegd op 10 november 2007).

[16] E. Rusty, *Processing XML with Java (tm): a guide to SAX, DOM, JDOM, JAXP, and TrAX* (Addison-Wesley 2002).

[17] E. Rusty, *XML in a nutshell. A desktop Quick Reference* (O'Reilly 2002).

[18] H.M. Blanken, *Intelligent search on XML data: applications, languages, models, implementations, and benchmarks* (Springer 2003).

[19] E.P. Lim, *Digital Libraries: peaople, knowledge, and technology* (Proceedings of the 5th international conference on Asian Digital Libraries, ICADL 2002, Singapore)(Springer 2002).

[20] E. Browne, *The myth of self-describing XML* (2003), http://www.oceaninformatics.biz/publications/e2.pdf (geraadpleegd op 10 november 2007).

[21] Zie voor een overzicht: http://en.wikipedia.org/wiki/List_of_document_markup_languages. Voor een vergelijking: http://en.wikipedia.org/wiki/Comparison_of_document_markup_languages (beide geraadpleegd op 10 november 2007).

[22] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485 (geraadpleegd op 10 november 2007).

[23] http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office (geraadpleegd op 10 november 2007). Zie hier ook voor

tion for the Advancement of Structured Information Standards (OASIS)[24]. Het bestandsformaat is gebaseerd op het XML-formaat zoals dat gebruikt werd door de Office-suite van OpenOffice.org. De ODF-standaard is een open standaard, wat betekent dat de specificatie gratis beschikbaar en implementeerbaar is. De gebruikte bestandsaanduidingen zijn .odt voor tekstdocumenten, .ods voor spreadsheets, .odp voor presentaties en .odg voor graphics. De eerste twee letters staan voor OpenDocument, de laatste letter geeft aan wat de specifieke toepassing is.
Een ODF-bestand is een gecomprimeerd JAR, Java Archive-bestand[25], dat bestaat uit een aantal bestanden en directories. Ongecomprimeerd worden de meeste gegevens vervat in eenvoudige tekstgebaseerde XML-files, zodat de data-inhoud de kenmerkende eenvoud van aanpassing, redactie en verwerking vertonen van alle XML-bestanden. De 'directories' worden gebruikt voor de opslag van documentbestanden, zoals images zonder SVG-formaat, animaties zonder SMIL (Synchronized Multimedia Integration Language[26])-formaat of andere, niet in XML uit te drukken formaten.
De JAR bestaat uit de volgende bestanden en directories:

1. XML-files
   - content.xml: het belangrijkste bestand, met de actuele content van het document (met uitzondering van de binaire data, zoals images). Het bestand is relatief eenvoudig leesbaar voor de mens.
2. Meta.xml: de metagegevens van het bestand, gestructureerd volgens de Dublin Core XML standaard[27].
3. Settings.xml: bevat instellingen zoals de zoomfactor of de positie van de cursor. Het betreft instellingen die niet door de content of de layout worden bepaald.
4. Styles.xml: de sjablonen of stijlen die gebruikt worden. ODF maakt hiervan uitgebreid gebruik, voor paragrafen, pagina's, fonts, frames, lijsten e.d.
5. Andere bestanden:
   - MIMEtype: de identificatie van het bestand die aangeeft wat voor soort bestand het is. Hierdoor is de bestandsaanduiding (.odt e.d.) alleen maar ten gunste van de gebruiker, die daaraan kan zien wat voor soort bestand het is.
6. Directories:
   - Meta/Inf: bevat meta-informatie over de binaire bestanden opgenomen in het JAR;
   - Thumbnails/: de binaire bestanden zelf.

ODF biedt een strikte scheiding tussen content, layout en metadata. De bestanden in XML-formaat worden verder gedefinieerd door een XML-schema gebaseerd op RELAX NG (REgular Language for XML Next Generation)[28]. RELAX NG is gedefinieerd door een OASIS-specificatie[29] en door deel 2 van de internationale standaard ISO/IEC 19757: Document Schema Definition Languages (DSDL)[30].
ODF wordt door een hele reeks applicaties ondersteund, waaronder OpenOffice.org, StarOffice, KOffice, IBM Symphony, GoogleDocs en andere[31]. Deze omgevingen ondersteunen ODF veelal als een 'native' bestandsformaat. Microsoft Office ondersteunt ODF middels een plug-in.
ODF maakt gebruik van de XML-recommendations van het World Wide Web Consortium (W3C)

---

de ODF-specificatie.

[24] http://www.oasis-open.org/home/index.php (geraadpleegd op 10 november 2007).

[25] Zie voor JAR: http://java.sun.com/j2se/1.3/docs/guide/jar/jar.html (geraadpleegd op 10 november 2007). Een JAR-file is qua structuur hetzelfde als een ZIP-bestand, maar heeft een directorystructuur met een vast gedefinieerde indeling. Doordat een JAR bestand ook een ZIP-file is, kunnen programma's die dit formaat kunnen weergeven ook worden gebruikt voor het weergeven van de inhoud van een JAR-bestand. Naast de Java-bestanden bevat het een extra-bestand, Metainf/Manifest.mf, waarin wordt aangegeven hoe het JAR-bestand gebruikt moet worden.

[26] Voor SMIL zie: http://www.w3.org/AudioVideo/ (geraadpleegd op 10 november 2007).

[27] ANSI/NISO Standaard Z39.85-2007. Zie: http://www.niso.org/standards/standard_detail.cfm?std_id=725 (geraadpleegd op 10 november 2007).

[28] Eric van der Vlist, *RELAX NG* (Cambridge 2003). Zie ook: http://relaxng.org/ (geraadpleegd op 10 november 2007).

[29] Zie: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=relax-ng (geraadpleegd op 10 november 2007).

[30] http://www.y12.doe.gov/sgml/sc34/document/0362_files/relaxng-is.pdf (geraadpleegd op 10 november 2007).

[31] Voor de hele lijst met ondersteunende applicaties zie: http://opendocumentfellowship.org/applications (geraadpleegd op 10 november 2007). In deze lijst zijn ook de plug-ins en filters opgenomen die het mogelijk maken ODF te gebruiken middels Microsoft Office-software.

en sluit zich zich aan bij de algemeen geaccepteerde XML-standaarden.
Veel van de op de ODF-standaard uitgebracht kritiek wordt in de nieuwe versie van de standaard (1.2., waaraan druk wordt gewerkt om nog in 2007 of begin 2008 te worden afgerond[32]) opgevangen.
De voornaamste kritiek op de ISO-specificatie van ODF is:

1. de keuze voor de W3C-'recommendation' MathML is volgens vele wiskundigen geen goede keuze. MathML is bestemd 'for the inclusion of mathematical expressions in Web pages' en 'machine-to-machine communication'. Door wiskundigen wordt veelal het TeX-formaat[33] gebruikt als methode voor het weergeven van complexe mathematische formules. TeX is geen ISO-standaard, maar is volledig gedocumenteerd. Daarnaast bestaat er kritiek vanwege het niet gebruiken van ISO 12083: 1994[34] voor mathematische formules. Deze ISO-standaard wordt niet ondersteund door MathML. In de nieuwe versie van de ODF-standaard 1.2. wordt niet van MathML afgeweken.
2. De specificatie bevat geen gedefinieerde formuletaal, waardoor problemen ontstaan aangaande de compatibiliteit[35]. OASIS ontwikkeld een standaard formuletaal met OpenFormula[36], welke zal worden opgenomen in de ODF-standaard 1.2..
3. De specificatie staat het gebruik van tabellen in presentaties niet toe[37]. Het gebruik ervan wordt gedefinieerd in de ODF-standaard 1.2. Over het algemeen wordt dit probleem nu ondervangen door een spreadsheet in de presentatie op te nemen en op die manier de benodigde functionaliteit te realiseren. Uiteraard is dit een onvolmaakte oplossing, omdat een spreadsheet- en een databasetabel fundamenteel verschillen.
4. De specificatie kent geen macro- of scripting-mogelijkheden, waardoor de verschillende applicaties op een afwijkende manier met deze functionaliteiten zullen omgaan. Hierdoor wordt de compatibiliteit tussen de verschillende ODF-gebruikende applicaties bedreigd[38]. Er wordt nog steeds gediscussieerd over de noodzaak om dergelijke mogelijkheden te definiëren, zodat opname in de ODF-standaard 1.2. niet voor de hand ligt.
5. De specificatie kent geen definitie van een digitale handtekening en beschouwt het als een applicatie-specifieke functionaliteit. ODF-standaard 1.2. zal de digitale handtekening opnemen volgens XML-Dsig[39].
6. De specificatie verwijst naar 'zip'-files zonder deze te refereren aan een 'standaard' die het zip-bestandsformaat beschrijft. Een dergelijke definiëring is echter beschikbaar via PKZIP[40] zelf, in de APPNOTE.TXT. Deze wordt continue onderhouden[41].

Office Open XML (OOXML)

OOXML is een op XML gebaseerd bestandsformaat voor electronische Office-documenten, zoals spreadsheets, grafieken, presentaties en tekstverwerkingsdocumenten. Het bestandsformaat is ontwikkeld door Microsoft, in oorsprong als een opvolger voor de binaire bestandsformaten van Microsoft Office.

[32] De concepten: http://www.oasis-open.org/committees/documents.php?wg_abbrev=office (geraadpleegd op 10 november 2007). Versie 1.1. van de standaard werd door OASIS in februari 2007 vastgesteld. De OASIS-standaarden zullen uiteindelijk leiden tot een nieuwe versie van de ISO/IEC 26300.
[33] Donald E. Knuth. *The TeXbook* (Computers and Typesetting, Volume A) (Reading (Ms.) 1984). Voor meer informatie: http://www.tug.org/ (geraadpleegd op 10 november 2007).
[34] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=20866 (geraadpleegd op 10 november 2007).
[35] Marco Fioretti, 'OpenDocument office suites lack formula compatibility', http://www.linux.com/articles/47942 (geraadpleegd op 10 november 2007).
[36] http://wiki.oasis-open.org/office/About_OpenFormula (geraadpleegd op 10 november 2007).
[37] Brian Jones., 'Quick question for ODF experts', http://blogs.msdn.com/brian_jones/archive/2006/07/20/673323.aspx (geraadpleegd op 10 november 2007).
[38] Marco Fioretti, 'Macros an obstacle to office suite compatibility', http://www.linux.com/articles/47935 (geraadpleegd op 10 november 2007).
[39] http://lists.oasis-open.org/archives/office/200702/msg00085.html . Voor XML-Dsig zie: http://www.w3.org/Signature/ (beide geraadpleegd op 10 november 2007).
[40] Voor PKZip: http://www.pkware.com/ (geraadpleegd op 10 november 2007).
[41] http://www.pkware.com/documents/casestudies/APPNOTE.TXT (geraadpleegd op 10 november 2007).

De specificatie werd ingebracht bij ECMA International ter standaardisatie[42]. De specificatie werd, na een betrekkelijk gesloten procedure door het ECMA Technical Committee (TC) 45, in december 2006 vastgesteld als ECMA-standaard 376[43]. De procedure behoefde niet 'open' te zijn, aangezien als uitgangspunt door TC 45 werd gesteld, dat de werkzaamheden van de commissie de productie betroffen van 'a standard which is fully compatible with the Office Open XML Formats, including full and comprehensive documentation of those formats in the style of an international standard, with particular attention given to enabling the implementation of the Office Open XML Formats by a wide set of tools and platforms in order to foster interoperability across office productivity applications and with line-of-business systems'[44]. Met kleine wijzigingen werd het door Microsoft ingediende voorstel overgenomen. ECMA heeft de vastgestelde ECMA 376[45] ingediend bij ISO/IEC ter standaardisatie. Het ingediende voorstel staat bekend als *ISO/IEC DIS 29500. Information technology -- Office Open XML file formats*[46]. Dit voorstel slaagde er niet in om voldoende ondersteuning te verwerven van de ISO-lidstaten in een procedure die tot versnelde standaardisatie moest leiden. Het voorstel zal opnieuw ter stemming worden gebracht en, indien voldoende aan de wensen van de lidstaten tegemoet wordt gekomen, kan alsnog versnelde standaardisatie plaatsvinden. Zo niet, dan wacht de lange procedure, die enkele jaren kan vergen.

Voorafgaande aan de standaardisatie van OOXML door ECMA, gebruikte Microsoft Office 2003 een enkel bestand, waarin in het XML alle binaire bestanden werden opgenomen[47]. Dat formaat wordt nu niet meer ondersteund. OOXML maakt gebruik van de Open Packaging Convention[48]. OPC is ontwikkeld door Microsoft voor de opslag van een combinatie van XML en niet-XML-bestanden die gezamenlijk één enkel document vormen in een enkele gecomprimeerde bestandscontainer. OPC vormt eigenlijk een profiel van het meer gewone ZIP-formaat. Als zodanig kan OPC niet alleen XML-gegevens en documentenbestanden bevatten, maar ook andere tekst en binaire bestandsformaten als PNG, BMP, AVI, PDF, RTF of zelfs een op JAR gebaseerd ODF-formaat. OPC specificeert de details van het ZIP-formaat, aangezien dit formaat niet als formele internationale standaard is gedefinieerd. Deze detaillering wijkt op onderdelen af van het bestand APPNOTE.TXT, zoals dat middels PKZip is verspreid, maar dit verhindert het openen van het Zip-bestand niet.

OPC gebruikt metadata aangaande relaties in plaats van directory- en bestandsnamen om individuele bestanden te localiseren. In de terminologie van OPC is een bestand een *part*. Een *part* heeft eveneens metadata die het vergezellen, vooral metadata aangaande MIME.

---

[42] http://www.ecma-international.org/memento/TC45.htm (geraadpleegd op 10 november 2007).

[43] Het persbericht: 'ECMA International approves Office Open XML Standard', http://www.ecma-international.org/news/PressReleases/PR_TC45_Dec2006.htm (geraadpleegd op 10 november 2007).

[44] http://www.ecma-international.org/memento/TC45.htm (geraadpleegd op 10 november 2007).

[45] Voor de specificatie: http://www.ecma-international.org/publications/standards/Ecma-376.htm (geraadpleegd 10 november 2007). De specificatie alleen telt ongeveer 6.000 pagina's, een ongekend hoog aantal. Zoals is opgemerkt, ligt het tempo van de vaststelling van deze specificatie door ECMA ver boven de gemiddelde door ECMA benodigde tijd voor het vaststellen van een standaard. 'Thus the remarkable achievement of Microsoft and Ecma TC45, who not only managed to create a standard an order of magnitude larger than any other markup standard I've seen, but at the same time managed to complete the review/edit/ approve cycle faster than any other markup standard I've seen. They have achieved an unprecedented review/edit/approval rate of 18.3 pages/day, 20-times faster than industry practice, a record which will likely stand unchallenged for the ages', Rob Weir, 'A notable achievement, *An antic disposition*, 9 december 2006, http://www.robweir.com/blog/2006/12/notable-achievement.html (geraadpleegd op 10 november 2007).

[46] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45515 (geraadpleegd op 10 november 2007).

[47] E.P. Tamminga, 'Office 2003 en XML. De gebruiker aan de XML!', http://download.microsoft.com/download/6/1/3/613785d4-409b-468f-9f98-68e4f11475fd/p39-43_1.02.pdf (geraadpleegd op 10 november 2007).

[48] OPC is gespecificeerd in Deel 2 van ECMA 376, http://www.ecma-international.org/publications/standards/Ecma-376.htm (geraadpleegd op 10 november 2007).

Een OPC bevat een XML-bestand, [Content_Types].xml in de root, en verder drie directories: _rels, docProps en een directory specifiek bestemd voor het document type dat gebruikt wordt (bijvoorbeeld: in een OPC met een tekstverwerkingsdocument, is er een Word-directory). Deze laatste directory bevat het document.xml-bestand, dat de inhoud bevat van het document. De afzonderlijke directories:

1. [Content_Types].xml: dit bestand beschrijft de inhoud van het OPC. Het bevat eveneens een overzicht van de bestandsformaten.
2. _rels: deze directory bevat de relaties inzake de bestanden in de OPC. Voor ieder specifiek bestand wordt een verwante _rels-directory aangemaakt, met daarin het bestand met de oorspronkelijke bestandsnaam, aangevuld met de toevoeging .rels.
3. _rels/.rel: in deze directory worden de relaties van het package zelf geplaatst. Applicaties die het OPC gebruiken zullen eerst hierin kijken.
4. word/document.xml: hier wordt het inhoudelijke deel van ieder tekstverwerkings (Word)-document opgeslagen.

OOXML wordt door geen enkele Office-suite als *native* formaat gebruikt, zelfs niet door Microsoft Office.

Het XML-formaat van Microsoft Office is gebaseerd op het (theoretische) OOXML van ECMA 376, maar is daarmee niet in overeenstemming. Aan het formaat zijn binaire elementen toegevoegd, macro's, OLE-objecten, Active X, Sharepoint metadata e.d.. Dit leidt tot incompatibiliteit met OOXML. Die incompatibiliteit wordt zowel bevestigd als ontkend, een tegenstrijdigheid die in een formele standaard niet mag voorkomen. Microsoft Office gebruikt namelijk bij het genereren van (nieuwe) documenten op uitgebreide schaal VML (Vector Markup Language). Dat is niet in overeenstemming met OOXML, tenminste als deel 4, sectie 6.1. moet worden geloofd. Daarin wordt aangegeven dat het formaat in principe enkel opgenomen is om conversie van de oude binaire Microsoft-formaten mogelijk te maken[49]. Uit deel 3, sectie 2.3.1. echter blijkt dat ook bij nieuwe documenten VML gebruikt wordt: 'All background information in a WordprocessingML document is stored using the Vector Markup Language (VML) syntax. The single exception to this is the background color, which is stored natively in WordprocessingML using the bgColor attribute'[50]. Het gebruik van VML negeert de industriestandaard SVG; 600 pagina's specificatie besteden aan het (slechts in hoofdlijnen) definiëren van een formaat dat in 1998 al door W3C als standaard is afgewezen, lijkt wat 'overdone'[51].

Het feit dat er geen overeenstemming bestaat tussen OOXML en het formaat dat in Microsoft Office is geimplementeerd is door Microsoft niet breed bekend gemaakt. OOXML kan evenwel zonder problemen in Microsoft Office worden geïmporteerd. Bij monde van Brian Jones, een van Microsoft's autoriteiten op het vlak van Microsoft's XML-formaten, weigert zijn bedrijf om officieel te verklaren dat het de toekomstige ontwikkeling en verbetering van de ECMA-standaard zal ondersteunen en dat ook die versie een gelijksoortige 'open' licentie[52] zal krijgen[53]. Het feit dat Microsoft niet garandeert (en eigenlijk zelfs weigert) de standaard als *native* formaat te gebruiken heeft aanleiding gegeven het bedrijf van cynisme en 'betrayal of trust' te beschuldigen[54]

[49] OOXML Specifications, Part 4 (Markup Language Reference), section 6.1., blz. 4343: 'The VML format is a legacy format originally introduced with Office 2000 and is included and fully defined in this Standard **for backwards compatibility reasons**. The DrawingML format is a newer and richer format created with the goal of eventually replacing any uses of VML in the Office Open XML formats. **VML should be considered a deprecated format included in Office Open XML for legacy reasons only** and new applications that need a file format for drawings are strongly encouraged to use preferentially DrawingML'. (Vetdruk toegevoegd door G.J. van Bussel). Zie: http://www.ecma-international.org/publications/standards/Ecma-376.htm (geraadpleegd op 10 november 2007).

[50] OOXML Specifications, Part 3 (Primer), section 2.3.1., blz. blz. 3-4. Zie ook: http://www.openmalaysiablog.com/2007/06/is-vml-in-or-ou.html (geraadpleegd op 10 november 2007).

[51] http://www.developer.com/xml/article.php/793961 (geraadpleegd op 10 november 2007).

[52] De discussie over de 'openheid' van de licentie is hevig (geweest), maar wordt hier niet verder uitgewerkt. We verwijzen daarvoor naar: http://www.grokdoc.net/index.php/EOOXML_objections (geraadpleegd op 10 november 2007).

[53] http://blogs.msdn.com/brian_jones/archive/2007/07/12/spreadsheet-formula-bugs.aspx (geraadpleegd op 10 november 2007).

[54] Frank Hayes, 'Microsoft won't commit to the open document standard it's pushing so hard', http://www.techworld.com/storage/features/index.cfm?f

Alle ondersteunende software maakt gebruik van een plug-in. Corel, Thinkfree Office, Gnumeric, Novell en NeoOffice zijn de voornaamste applicaties waarin OOXML op meer of mindere mate lees- en schrijfbaar is. Al deze applicaties nemen het OOXML van ECMA 376 als uitgangspunt. ECMA 376 fungeert voor Microsoft Office eigenlijk alleen als een importformaat.
Er zijn uitzonderlijk veel bezwaren aangetekend tegen de specificatie van OOXML. Een gedetailleerd overzicht van alle kritiek is door de website www.grokdoc.net samengesteld[55]. Wij beperken het overzicht tot de meest fundamentele opmerkingen:

1. de specificatie negeert of gaat in tegen diverse internationale standaarden, zoals de Gregoriaanse kalender, de ISO 8601: 2004 (over de numerieke representatie van datum en tijd)[56], de ISO 639 -1:2002 (codes voor de representatie van namen en talen)[57], de ISO/IEC 8632 : 1999 (Computer Graphics Metafile)[58], de ISO/IEC 26300: 2006 (Open Document Format for Office Applications), W3C SVG (Scalable Vector Graphics), W3C MathML, de ISO/IEC 10118-3 (hash-functies)[59], W3C XML-ENC (encryptie)[60] en W3C SMIL.
2. de specificatie is niet voldragen en niet consistent, blijkende uit onder andere: niet-bestaande of niet-gebruikte maateenheden, verwarrende en inconsistente definiëringen van hexadecimale getallen, niet-gespecificeerde begrippen en termen, inconsistente benamingen voor elementen en attributen, niet-XML coderingen, herdefinitie van standaard kleurenwaarden en het gebruik van Microsoft-specifieke waarden en standaarden (zoals bijvoorbeeld VML en DrawingML). Meer dan 10 % van de voorbeelden gebruikt in het tekstverwerkingsdeel van de specificatie kunnen niet worden gevalideerd op basis van het gebruikte schema, iets wat tijdens het standaardisatieproces bij ECMA had moeten worden hetsteld.
3. de specificatie bevat een groot aantal bitmasks, waardoor interoperabiliteit en leesbaarheid door de mens (een van de grote voordelen van XML) wordt beperkt. Een bitmask is een techniek om verschillende waarden te coderen in een enkele variabele, door betekenis toe te kennen aan iedere individuele bit van de variabele. Als voorbeeld: het binaire 10110001 (177) betekent als bitmask: Ja/Nee/Ja/Ja/Nee/Nee/Nee/Ja, en geeft antwoord op acht verschillende ja/nee-vragen. In OOXML komen in ieder geval bitmasks voor inzake 'paragraph conditional formatting, table cell conditional formatting, table row conditional formatting, table style conditional formatting settings exceptions', etc[61]. Bitmasks creëren eigenlijk een nieuw datamodel, afgescheiden van XML. Een bitmask kan niet worden beschreven of gevalideerd door XML-schema's als RelaxNG of Schematron[62]. Bitmasks maken ook het werken met XSLT[63], een standaard van W3C voor manipulatie en conversie van XML-documenten, onmogelijk. XSLT heeft geen tools voor het werken met bitmasks, aangezien deze geen deel uitmaken van het XML-gegevensmodel.

eatureid=3685&pagtype=all (geraadpleegd op 10 november 2007).
[55] 'EOOXML objections', http://www.grokdoc.net/index.php/EOOXML_objections (geraadpleegd op 10 november 2007).
[56] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=40874 (geraadpleegd op 10 november 2007).
[57] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=22109 (geraadpleegd op 10 november 2007).
[58] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32378 (geraadpleegd op 10 november 2007).
[59] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=37416 (geraadpleegd op 10 november 2007).
[60] http://www.w3.org/TR/xmlenc-core/ (geraadpleegd op 10 november 2007).

[61] Rob Weir, 'A bit about the bits with the bits', *An Antic Disposition*, 10 november 2006, http://www.robweir.com/blog/2006/10/bit-about-bit-with-bits.html (geraadpleegd op 10 november 2007). De genoemde bitmasks zijn te vinden resp.: OOXML Specifications, Part 4 (Markup Language Reference), section 2.3.1.8., blz. 59-60, section 2.4.7.,blz. 302-303; section 2.4.7., blz. 303-305 section 2.4.52, blz. 430-431. Zie: http://www.ecma-international.org/publications/standards/Ecma-376.htm (geraadpleegd op 10 november 2007).
[62] Voor Schematron: http://www.schematron.com/ (geraadpleegd op 10 november 2007).
[63] Voor XSLT: http://www.w3.org/TR/xslt (geraadpleegd op 23 oktober 2007). Zie ook: John R. Gardner, Zarella L. Rendon, *XSLT & XPath, A Guide to XML Transformations* (New York 2001).

4. de specificatie berust voor een deel op niet-vrijgegeven informatie, die, indien door anderen gerealiseerd, leidt tot inbreuken op (niet onder de door Microsoft gehanteerde gebruikerslicentie vrijgegeven) patenten van Microsoft[64]. De specificatie kan derhalve niet (of nauwelijks door andere leveranciers worden geimplementeerd:
    1. ze vereist emulaties van eerdere Microsoft producten, waarvan het gedrag verder niet wordt gespecificeerd. Aangezien deze producten 'proprietary software' zijn kan alleen Microsoft deze delen van de specificatie implementeren;
    2. ze vereist implementatie van Windows Metafile[65] in plaats van ISO/IEC 8632: 1999[66]. Dit is 'proprietary' technologie, die in de specificatie niet verder wordt uitgewerkt. Windows Metafile is veelvuldig gebruikt door andere leveranciers voor implementaties. De door Microsoft uitgegeven licentie geldt niet voor niet-vrijgegeven of geopenbaarde informatie. In de specificatie wordt Windows Metafile niet nader gedefinieerd.
5. de specificatie verwijst naar 'application-defined' gedrag om belangrijke functionaliteit te ondersteunen, die *an sich* gedocumenteerd had moeten zijn of ondersteund door bestaande standaarden. Ook hierdoor wordt interoperabiliteit ernstig belemmerd. Enkele voorbeelden:
    1. het 'equationxml'-attribuut van 'shape'-elementen wordt beschreven, 'used to rehydrate an equation using the Office Open XML Math syntax'. Dit is bedoeld voor de essentiële functionaliteit de wiskundige vergelijkingen in tekeningen te kunnen redigeren en interpreteren op basis van de onderliggende wiskundige structuur. Maar, zo gaat de specificatie verder, 'the actual format of the contents of the attribute are application-defined', waardoor uitwisseling tussen applicaties onmogelijk wordt[67].
    2. het 'gfxdata'-attribuut van 'shape'-elementen wordt beschreven, welke 'contains DrawingML content' die 'base-64 encoded' is. Maar de 'contents of this package are application-defined', dus ook al gebruiken ze 'the Parts defined by this Standard whenever possible' is er niet voldoende informatie voor een onafhankelijke implementatie om de 'DrawingML content' te kunnen lezen of afbeelden[68].
    3. er wordt een 'ink'-element beschreven, die 'ink-annotations' opslaat 'in an application-defined format'. Dit is zonder twijfel bedoeld voor de opslag van Microsoft Ink-aantekeningen, waarmee handgeschreven notities aan documenten kunnen worden toegevoegd[69]. Deze 'Ink'-notaties zijn niet gedefinieerd in de specificatie, waardoor uitwisseling van deze gegevens tussen applicaties zo goed als onmogelijk is. De noodzaak om 'Ink' te gebruiken is niet aanwezig: in bestaande standaarden als W3C PNG[70] en W3C SVG zijn voldoende mogelijkheden om de betreffende bestanden weer te geven.
6. de specificatie heeft, volgens het Adaptive Technology Resource Center van de Universiteit van Toronto, 'grave issues with respect to the accessibility ... as a format and potential standard that should preclude its adoption at

[64] Matthew Cruickshank, Chris Daish, Conal Tuohy, 'Microsoft and Open Standards. Can Other Vendors Implement Microsoft's Office Open XML?', 15 augustus 2007, http://holloway.co.nz/can-other-vendors-implement-ooxml.html (geraadpleegd op 10 november 2007).

[65] Voor Windows Metafile: http://www.fileformat.info/format/wmf/ (geraadpleegd op 10 november 2007). Zie ook: 'Windows Metafiles. A guide for non-windows programmers', http://www.skynet.ie/~caolan/publink/libwmf/libwmf/doc/index.html (geraadpleegd op 10 november 2007).

[66] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=32378 (geraadpleegd op 10 november 2007).

[67] OOXML Specifications, Part 4 (Markup Language Reference), section 6.1.2.19., blz. 4653.

[68] OOXML Specifications, Part 4 (Markup Language Reference), section 6.1.2.19., blz. 4655.

[69] OOXML Specifications, Part 4 (Markup Language Reference), section 6.2.2.14., blz. 4813-4814. Voor Ink: http://www.microsoft.com/windowsxp/using/tabletpc/learnmore/vanwest_03may28inkword.mspx (geraadpleegd op 10 november 2007).

[70] Voor PNG: http://www.w3.org/TR/PNG/ (geraadpleegd op 10 november 2007).

present. It may be the case that OOXML can be improved to ameliorate some of the more specific technical concerns, but it is most likely too late for the higher-level issues, especially those inherent in the process by which OOXML was developed. We suggest that energy would be better spent in the ongoing effort to improve the existing ISO ODF standard (with which OOXML would overlap and compete if it is adopted). In any event, decisions with respect to standardized document formats should be made in consultation with members of disability communities, disabilities experts and developers of assistive technologies, with universal accessibility as a core requirement as opposed to an *ad hoc* afterthought'[71].

V. De validiteit van de argumenten voor het gebruik van XML-bestandsformaten

Vooraf dient hier te worden aangetekend dat alle voor- en nadelen die aan het gebruik van XML worden toegekend (en die hiervoor zijn benoemd) van toepassing zijn (verklaard) op de XML-bestandsformaten. Daarnaast zijn er voor de acceptatie van XML-bestandsformaten, met name in overheidskringen, drie argumenten aangevoerd, te weten:

1. XML-formaten vormen een open standaard;
2. XML-formaten bevorderen de interoperabiliteit tussen applicaties en platformen;
3. XML-formaten zijn duurzaam.

Deze argumenten worden niet zozeer aangevoerd op grond van de *merites* van een inhoudelijke analyse van de betreffende XML-bestandsformaten, maar veeleer omdat XML zelf deze argumenten ondersteunt.

XML-formaten vormen een open standaard.

Een open standaard is van niemand eigendom en mag door iedereen gebruikt worden. De Nederlandse overheid definiëert de volgende eisen waaraan open standaarden dienen te voldoen:

1. vaststelling vindt plaats op basis van een open beslissingsprocedure, waarbij sprake is van consensus of meerderheidsbeslissing;
2. beheer van de standaard is toegewezen aan een not-for-profit organisatie met een vrij toetredingsbeleid;
3. de standaarden zijn gepubliceerd;
4. het gebruik van de standaard kent lage kosten, opdat geen drempel voor toegang tot de standaard bestaat;
5. intellectueel eigendom dat aan de standaard ten grondslag ligt wordt zonder kosten ter beschikking gesteld aan gebruikers en implementatoren;
6. er zijn geen beperkende voorwaarden voor het (her-)gebruik van de standaard;
7. er wordt geen afhankelijkheid gecreëerd van een specifiek bedrijf of organisatie;
8. een standaard is gratis te implementeren;
9. vanuit de standaardorganisatie wordt geen voorkeur gegeven aan een leverancier[72].

Dit eisenpakket is volledig van toepassing op ODF, zelfs al zou het enkel een OASIS-standaard zijn. Vaststelling van versie 1.2. van ODF door OASIS maakt deze versie, ook al heeft vaststelling door ISO daarvan nog niet plaatsgehad, eveneens tot een open standaard. Aan alle negen hiervoor genoemde eisen kan ODF, in alle versies, voldoen. Het eisenpakket is voor een deel niet toepasbaar op OOXML. Het bestandsformaat botst (in dit stadium) minimaal met de eisen 1, 5 en 7. Het standaardisatieproces bij ECMA was (en is) een gesloten proces. Deze eis wordt ingevuld indien de IEC-ISO besluit OOXML als ISO-norm vast te stellen, waarbij aangetekend dient te worden dat verdere ontwikkeling van de specificatie van OOXML geopend moet worden voor andere partijen om werkelijke 'openheid' te realiseren. Er bestaat (voor wat betreft eis 5) geen probleem met het intellectueel eigendom dat is beschreven in de specificatie. Niet-vrijgegeven informatie echter is vereist voor het op de juiste wijze kunnen implementeren van het bestandsformaat, wat welhaast zeker leidt tot het moeten gebruiken van niet-vrijgegeven patenten. Daaronder valt bijvoorbeeld de Windows Metafile. De niet-vrijgegeven informatie leidt ook tot een afhankelijkheid van Microsoft om tot een volwaardige implementatie te kunnen komen.

[71] Stephen A. Hockema, Jutta Treviranus, 'Accessibility Issues with Office Open XML', http://atrc.utoronto.ca/index.php?option=com_content&sectionid=14&task=view&hidemainmenu=1&id=3 (geraadpleegd op 10 november 2007).

[72] Voor de eisen 1 tot en met 6 zie: http://www.ososs.nl/wat_zijn_open_standaarden; voor de eisen 7 tot en met 9 zie: http://webrichtlijnen.overheid.nl/handleiding/ontwikkeling/productie/open-standaarden/ (beide geraadpleegd op 10 november 2007).

De aanduiding 'open standaard' volgens het bovengenoemde uitgangspunt is met recht van toepassing op het XML-bestandsformaat ODF. OOXML kan volgens datzelfde uitgangspunt echter (nog) geen open standaard genoemd worden.

XML-formaten bevorderen de interoperabiliteit tussen applicaties en platformen

Interoperabiliteit kan omschreven worden als 'the ability of a system or a product to work with other systems or products without special effort on the part of the customer'[73]. Het is helder dat wanneer een XML-schema *volledig* gespecificeerd is en akkoord gegeven door alle partijen, het nuttig zal zijn voor interoperabiliteit. Wat vaak wordt vergeten is dat ook al gaan alle partijen akkoord, dat nog niet wil zeggen dat alle partijen de mogelijkheid hebben om de gegevens in het XML-schema te maken of te gebruiken. Ook in dit geval ontstaat er een probleem met interoperabiliteit.
Bij beide bestandsformaten mag een gerede twijfel bestaan over de interoperabiliteit, bij ODF wat minder, bij OOXML wat meer.
Bij ODF wordt in de huidige versie de interoperabiliteit bedreigd door ontbrekende Office-functionaliteiten (zoals een formule-taal voor spreadsheets) en 'application-specified' gedrag voor macro's. Macro's vormen een dermate belangrijk onderdeel van de hedendaagse Office-applicaties, dat ze niet aan de verschillende applicaties kunnen worden overgelaten en in de specificatie moeten worden gedefinieerd. Ook in de nieuwe standaard-specificatie (1.2.) worden macro's niet betrokken.
Bij OOXML is interoperabiliteit een groter probleem. De interoperabiliteitskritiek op OOXML is tot de volgende argumenten terug te brengen:

1. veel informatie benodigd voor volledige interoperabiliteit is niet beschikbaar of vrijgegeven. Bitmasks en andere gegevens met een 'proprietary' karakter maken interpretatie en gebruik door andere partijen dan Microsoft moeilijk[74];
2. OOXML neigt ernaar bestaande standaarden en algeheel geaccepteerde richtlijnen te negeren, aan te passen naar goeddunken of te vervangen door eigen gedefinieerde standaarden, die voor andere partijen dan Microsoft nieuw zijn of niet ondersteund kunnen worden;
3. belangrijke functionaliteiten van Office-applicaties zijn niet beschreven, maar worden overgelaten aan 'application-specified' gedrag, zoals, net als bij ODF, macro's, maar ook andere functionaliteiten.

Interoperabiliteit is problematisch bij beide XML-gebaseerde bestandsformaten. ODF is het meest interoperabel, OOXML het minst. Deze laatste vertoont wel een uitmuntende interoperabiliteit met de oudere, binaire bestandsformaten van Microsoft. Aangezien er vanuit Microsoft geen geheim gemaakt wordt van het feit dat OOXML daarvoor ook bedoeld is, hoeft dat niet te verrassen[75]. Interoperabiliteit van OOXML als bestandsformaat blijft daarbij echter ver achter.

XML-formaten zijn duurzaam.

In theorie beschermt XML documenten tegen de afhankelijkheid van uitstervende besturingssystemen en daaronder gebruikte software door de onafhankelijkheid van platforms en programmatuur. De overtuiging bestaat dat een digitaal document geschreven in XML in de tijd duurzaam zal zijn[76]. Het probleem is dat duurzaamheid, dit wil zeggen het kunnen reconstrueren van documenten of objecten op elk moment in de tijd, met de vorm,

[73] http://searchsoa.techtarget.com/sDefinition/0,,sid26_gci212372,00.html (geraadpleegd op 10 november 2007). Ook: Paul Miller, 'Interoperability. What is it and Why should I want it?', http://www.ariadne.ac.uk/issue24/interoperability/intro.html (geraadpleegd op 10 november 2007).

[74] Computer & Communications Industry Association, 'Microsoft's approach to Disclosures of XML File Formats for Word 2003 and Excel 2003' (2004). Deze publicatie is gepubliceerd op de website van de organisatie (http://www.ccianet.org/papers/CCIA-XML.pdf), maar is daar verwijderd. Deels kan de argumentatie van de CCIA terugelezen worden op: http://www.groklaw.net/article.php?story=20051020193905892 (geraadpleegd op 10 november 2007).

[75] Zie onder andere: http://download.microsoft.com/download/7/4/3/7437e747-aeaf-4419-8181-7307bae89db4/XMLFileFormats-Guide.doc (geraadpleegd op 10 november 2007).

[76] Indien tenminste de XML-specificatie, de Unicode-tabel, schema's van de typen documenten met documentatie alsmede Engelse en Nederlandse woordenboeken (om de betekenis van de tag-namen te herleiden) bijgevoegd worden, zoals in de white paper *XML en digitale bewaring* van het Testbed Digitale Bewaring terecht wordt opgemerkt. Zie: http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white-paper_xml-nl.pdf (geraadpleegd op 10 november 2007).

structuur en inhoud van het moment waarop ze werden gemaakt of vastgesteld, geen prioriteit is bij de ontwikkeling van XML. Soft- en hardwareleveranciers hebben geen behoefte aan een duurzaam formaat, wel aan een formaat dat op alle platformen op dezelfde wijze kan worden gepresenteerd en gelezen (al betekent dat nog niet dat dat met alle software zou moeten kunnen)[77].
XML is nauwelijks een decennium oud. Het is nauwelijks beproefd daar waar het duurzaamheid betreft. En ieder XML-bestandsformaat heeft één essentieel duurzaamheidsprobleem: de authenticiteit en de integriteit van de documenten is niet gewaarborgd en zelfs op uiterst simpele wijze aan te tasten. XML is immers van nature muteerbaar; het is ook niet ontwikkeld om documentaire duurzaamheid te waarborgen. Uiteraard kunnen allerlei soft- en hardwarevoorzieningen aangewend worden om onmuteerbaarheid te waarborgen, maar die voorzieningen verdampen met de onontkoombare vervanging van hard- en software. Duurzaamheid van de inhoud is dan ook bij beide XML-bestandsformaten zeer twijfelachtig.

Conclusie

Als bestandsformaten voor Office-applicaties hebben ODF en OOXML gemeen dat ze beide gebaseerd zijn op XML. Daarmee vertonen ze alle voor- en nadelen van XML zoals die hiervoor beschreven zijn. De XML-schema's die aan beide bestandsformaten ten grondslag liggen vertonen problemen:

1. bij ODF omdat het schema te beperkt is, waardoor het bestandsformaat op onderdelen afhankelijk wordt van 'application-specified' gedrag voor het realiseren van functionaliteiten;
2. bij OOXML omdat het schema vertrouwt op niet-vrijgegeven informatie, bitmasks en (veelvuldig) op 'application-specified' gedrag voor het realiseren van functionaliteiten. Daarnaast negeert het schema bestaande, wijdverbreide open standaarden, om daar eigen interpretaties voor in de plaats te stellen.

Als het gaat een bestandsformaat te gebruiken dat open is en de meeste interoperabiliteit biedt dan heeft van de beide bestandsformaten ODF (zeker met de nieuwe versie 1.2. van de specificatie) de voorkeur, vooral ook omdat de interoperabiliteit met oudere binaire bestandsformaten acceptabel is.
Als het gaat een bestandsformaat te gebruiken dat duurzaam is voor wat betreft de inhoud van de documenten dan is een XML-bestandsformaat niet direct voor de hand liggend gezien het blijvend muteerbare karakter ervan. Daarbij: duurzaamheid is geregeld in een andere ISO/IEC-standaard, namelijk die van PDF/A (ISO/IEC 19005-1. Document management - Electronic document file format for long-term preservation)[78].
Beide bestandsformaten kunnen wel als een 'container' worden gebruikt, waarbij documenten in hun oorspronkelijke bestandsformaat (bijvoorbeeld PDF/A) kunnen worden ingekapseld. Hierdoor wordt het originele document opgenomen, zonder dat de inhoud en layout kunnen worden gewijzigd, waardoor de integriteit behouden blijft (er vindt namelijk geen verandering van de bitstream plaats). In het XML-deel van het bestand kunnen allerlei aanvullende metadata worden opgenomen, terwijl het uit het oogpunt van flexibiliteit ook mogelijk is meerdere representaties van het document op te nemen[79].

[77] Softwareleveranciers zijn van nature geneigd te kiezen voor een 'proprietary' bestandsformaat aangezien een dergelijk formaat hun eigen markt beschermd. Daarbij: duurzaamheid is commercieel minder interessant. Hardwareleveranciers hebben behoefte aan een korte levensduur en veel verschillende versies van soft- en hardware, met veel conversies en migraties, vanwege de positieve effecten daarvan op hun omzet. Emulaties en virtualisaties vragen daarnaast om veel geheugenruimte en hardware.

[78] Zie: www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920 (geraadpleegd op 10 november 2007).

[79] Ronald Dekker, Martin Slabbertje, 'Het duurzaam bewaren van wetenschappelijke digitale bronnen', *Informatieprofessional* 7 (2003), nr. 6, blz. 32-34. Zie ook: http://igitur-archive.library.uu.nl/DARLIN/2005-0526-200314/UUindex.html (geraadpleegd op 10 november 2007).