# TOWARD XML-BASED OFFICE DOCUMENTS
## *(A BRIEF INTRODUCTION)*

JACEK POLEWCZAK

Contents



## 1. What is happening?

Recent trends in office document formats indicate a move towards open and standard-based XML formats. Major government agencies and public and private institutions started looking for office documents formats that assure compatibility with open standards, that are vendor neutral, cross-platform interoperable, and non-binary (i.e., XML-based).

Although the pressure to embrace open file formats has been felt for many years, Microsoft, with the dominant role in office documents formats, has been reluctant to move from its proprietary, binary formats to open document standards. The situation only *seemingly* changed in November 2005 when Microsoft, with a number of industry partners and supporters, took steps to produce an open specification for their own Office file formats. In December 2006, the specification was approved by ECMA International (European Association for Standardizing Information and Communication systems) as ECMA-376: *Office Open File Formats* (OOXML). In April 2008, OOXML (as ISO/IEC DIS 29500) received necessary votes in the ISO (International Organization for Standardization) for approval as an international standard.

However, in developing OOXML, Microsoft have chosen not to support ISO/IEC 26300: *Open Document Format for Office Applications* (ODF), submitted to ISO by OASIS (Organization for the Advancement of Structured Information Standards), and approved by ISO as international standard in May 2006. The benefits of OOXML, Microsoft argued, are concentrated on backwards compatibility with its legacy binary file formats. (see Section 4).

Section 2 provides the details of requirements and needs government agencies and public and private institutions look for in their search for open document formats. In Sections 3 and 4, I provide some technical details and comparisons of *Open Document Format for Office Applications* (ODF) and *Office Open File Formats* (OOXML), respectively. Section 5 is about PDF file format. I conclude with the recommendations in Section 6.

*Date*: November 4, 2008.

## 2. Changing attitudes

Recently, many government agencies and public and private institutions were looking toward *A strategy for Openness*. This was the theme of the Report to the Governor and Legislature of New York State (May 2008). In its executive summary, the workgroup developed the following recommendations (page 10 of [4]) to ensure the State's electronic records are:

- Created and preserved in ways that encourage choice, interoperability, and vendor neutrality;
- Accessible to the public; and
- Kept under proposed appropriate government control.

In the state of Massachusetts (since 2005), the state agencies are required to create and save "official records" in one of the following "open" or "acceptable" formats:

- ODF (open)
- OOXML (open, added in 2007)
- HTML (open)
- ASCII (open)
- RTF (acceptable)
- PDF (acceptable)

The states of California, Connecticut, Florida, Minnesota, Oregon, and Texas introduced bills that asked to create, exchange, and preserve all documents in open file (preferably XML-based) formats (see page 6 of [4]). Except for the state of Minnesota, the bills still remain in committees.

The European Union (EU) and the United Nations (UN) have been interested for many years in the use of open standards to facilitate electronic transfer of information. In a comprehensive report (2003), the Valoris consulting group, contracted by EU, identified the following criteria by which competing office document formats could be judged (see [5]):

- use of open standards
- being non binary (i.e., XML-based)
- capable of being modified
- preserving format fidelity
- offering cross-platform interoperability
- supporting current word processor features
- supporting future word processing features
- being widely adopted

In December 2006, the Pan-European eGovernment Service Committee (PEGSCO) issued *Conclusions and Recommendations on Open Document Formats* to "public administrations" and "industry, industry consortia and international standardisation bodies" (see [7]):

> In view of the present situation, public administrations are invited:
>
> - To make maximal use of internationally standardized open document exchange and storage formats for internal and external communication;
> - To use only formats that can be handled by a variety of products, avoiding in this way to force the use of specific products on their correspondents. When the usage of proprietary formats is unavoidable, alternative, internationally standardized open formats shall be provided in addition to proprietary formats;
> - To adapt, where appropriate, national guidelines and regulations, taking into account the arrival of international standards in this area;

- To consider the definition of minimum requirements in regard to the functionalities of open document exchange formats in view of pursuing the compatibility of applications;
- To create guidelines for the use of revisable and non-revisable document exchange and storage formats for different purposes.

Industry, industry consortia and international standardisation bodies are invited:

- To work together towards one international open document standard, acceptable to all, for revisable and non-revisable documents respectively;
- To develop, applications that can handle all relevant international standards, leaving the choice to their customers as to what format will be used "by default";
- To avoid invalidating the purpose of open document exchange and storage formats by offering extensions to the relevant international standards as default formats;
- To make proposals for conformance testing and to develop adequate tools in order to safeguard interoperability between applications;
- To continue to improve the existing standards, also taking into account additional needs such as electronically signed documents.

Since 2006, numerous governments adopted policies of (often strict) adherence to open document formats. These are: Belgium, Brazil, Croatia, Denmark, France, Japan, Malaysia, Netherlands, Norway, Poland, Russia, South Africa, Switzerland, Uruguay. The list also includes regional/provincial/state governments. (See [6] for the current list.) Further details can be found in the fourth report ([8]) of the Center for Strategic and International Studies on the use of open source products by the government agencies.

The universities and higher education sector must also concentrate on open document formats. The following quote from Walter Ditch's comprehensive report ([9], 2007, Higher Education Funding Council for England) characterizes the situation in UK, although, it applies to the US higher education sector as well:

> *The report proposes that although the UK higher education sector has, for a long time, understood the interoperability benefits of open standards, it has been slow to translate this into easily understandable guidelines for implementation at the level of everyday applications such as office document formats. As far as higher education is concerned, the use of office document formats has now reached a watershed. There is an urgent need for co-ordinated, strategically informed action over the next five years, if the higher education community is to facilitate a cost effective approach to the switch to XML-based office document formats.*

## 3. ODF

Acquisition of a small German software company StarDivision in 1999 was Sun Microsystems' entry into the office application market, dominated by Microsoft Office. In contrast to Microsoft Office binary file formats, the StarOffice package, and subsequently OpenOffice.org, used XML for its file format. The StarOffice, intended for corporate users, is priced at approximately $70 USD, while OpenOffice.org is a free, open source office application. XML format used by OpenOffice v.1.0 was developed by OASIS (Organization for the Advancement of Structured Information Standards) into an open standard. In 2005, OASIS submitted it to ISO for ISO/IEC approval, and in May 2006, ODF was approved as the internationally recognized office document file format, ISO/IEC 26300:2006 *Open Document Format for Office Applications* (ODF).

Further development of ODF has been carried by OASIS. Details of all versions of the ODF specification are available from the ODF Technical Committee home page.

ODF is an XML-based file format that facilitates the creation and editing of documents containing text, spreadsheets, charts, presentations and graphics. The ODF specification reuses existing open standards, or portions of such standards, and thus, it reduces the complexity of the standard itself. These include XSL-FO, SVG (scaled vector graphics), XLink, XForms, MathML, and DublinCore. Its 700+ pages of specifications is contrasted with 6000+ pages of specifications of OOXML. A good overview of the ODF as well as its specification (ver.1.0) is available online, see [11] and [12].

There are already dozens of implementations and applications that support ODF on a variety of operating systems, including Linux, Mac and Windows platforms. An incomplete list include OpenOffice.org, Sun StarOffice 8, KOffice, IBM Lotus Notes 8 Documents, IBM Lotus Symphony Documents, Softmaker Office 2008, Apple TextEdit, AbiWord, Google Docs & Spreadsheets, Zoho Writer, AjaxWriter, and Corel WordPerfect. The complete list, including applications for text, spreadsheets, and presentation documents, as well as content management system applications of ODF, are listed in [13].

Following [9], I list important technical advantages and disadvantages of ODF.

Technical advantages of ODF

- Simple specifications, building on existing open standards;
- Supported by multiple applications on all platforms, including Linux, Mac, and Windows;
- ODF uses a mixed content markup model, with very good separation of content and presentation.

Technical disadvantages of ODF

- ODF is insufficiently detailed: Spreadsheet formulae are applications defined[1];
- Macro/scripting is not defined in ODF;
- No support for digital signatures[2].

## 4. OOXML

Similarly to ODF, the OOXML file format is based on a compressed Zip archive. In December 2006, OOXML was approved by ECMA International (European Association for Standardizing Information and Communication systems) as ECMA-376: *Office Open File Formats* (OOXML). In April 2008, OOXML (as ISO/IEC DIS 29500) received necessary votes in the ISO (International Organization for Standardization) for approval as an international standard. The detailed specification of OOXML file format can be found in [1]. OOXML uses three custom XML-based languages to describe types of document content: WordProcessingML, SpreadSheetML, and PresentationML.

At present time very few applications support OOXML. Furthermore, even Microsoft Office 2007 does not support OOXML, as defined in ISO/IEC DIS 29500 or ECMA-376. In addition to Microsoft Office 2007, only Microsoft Office 2008 for Mac OS X has native support for OOXML. Partial and not confirmed at this time support for OOXML has been announced by Novell's edition of Open Office and Corel's WordPerfect.

As in the case of ODF standard (see [9]), I list important technical advantages and disadvantages of OOXML.

[1]Detailed spreadsheet formula syntax, OpenFormula is already included, see [14]

[2] Support for digital signatures will be included in ODF ver.1.2, see [15]

Technical advantages of OOXML

- Backwards compatibility with existing Microsoft proprietary binary formats;
- Faster operation and better memory use.

Regarding the backwards compatibility of OOXML with Microsoft proprietary binary formats, the following comments are in order:

(1) From Google comments about OOXML ([16]):
> *. . . if OOXML were necessary to faithfully convert these legacy documents to an XML format, it would have to contain the complete specification of these older document formats. Without this OOXML would be incomplete in its descriptions for an ISO standard. No specifications for older document formats exist in the OOXML descriptions, and so any argument that OOXML is needed for their accurate translation is false. Such legacy documents may just as easily be translated to ODF (as can be seen in the way some existing ODF implementations handle the import of the legacy Microsoft Office file formats).*

(2) In reply to criticism (1) above, Microsoft posted on its site (see [17] and [18]), and thus outside the ISO scope, the binary Office document specifications. However, as the Oracle Corporate Architecture group had noticed (see [19]), NO standardized mapping of binary formats to OOXML were provided, and Microsoft refused to provide such mappings before the ballot took place on April 2, 2008. This meant that, except for Microsoft Office, no other application supporting OOXML would be able to faithfully recreate the look of Microsoft's legacy binary formats.

Technical disadvantages of OOXML (see [9])

- Inconsistencies with existing ISO standards:
  Examples of these include: Paper sizes (ISO 216 defines names for paper sizes, whereas OOXML uses its own numeric codes for these sizes); Date and Times are covered in ISO 8601, but OOXML makes use of an alternative mechanism which considers 1900 as a leap year and does not understand dates prior to 1900 (an existing error found in Microsoft Office legacy spreadsheets); HTML colour names (ISO/IEC 15445).
- Inconsistencies with existing W3C Recommendations:
  OOXML defines its own vector graphics markup (DrawingML) rather than making use of SVG. This may be in order to remain backwards compatible with an earlier Microsoft Office drawing format, VML. A counter argument to this criticism is that standards such as SVG may not be wholly suitable for the required purpose, leading to a requirement to invent a new solution, or to adapt a standard to an excessive degree. Support for this viewpoint comes from the unlikely source of Sun's own development community ([20]), However, in addition, OOXML does NOT make use of the W3 recommended mathematics markup language, MathML.
- Cloning behaviour of undocumented legacy features:
  Several sections of the OOXML specification make reference to behaviour of an application without defining the nature of that behaviour. For example, 'autoSpaceLikeWord95'. It is argued that only Microsoft can implement these proprietary features and therefore OOXML cannot be reasonably implemented by others (for an extended list of such features see [21]).
- Size of the specification:
  The OOXML standard specification has 6000+ pages and responses to the Ecma

International standardization process have argued that this is a serious issue which results from the failure to leverage existing, open standards within the standard.

- The use of a separate "relationships" file to hold hyperlinks:

  It has been argued that this may cause problems with the manipulation of the XML in an OOXML document and, in particular, may affect the use of he standard translation tool, XSLT. This needs to be clarified, as it is potentially very serious, since the inability to transform the XML would restrict the repurposing of the information contained in the file, and would also inhibit easy conversion to other formats (for example to html and pdf formats.
- Macro/scripting language is not defined in OOXML.
- Specification is incomplete:

  There are elements in the Microsoft's Office 2007 file formats that are not documented in Ecma-376 e.g., VBA. This may cause interoperability problems with applications that utilize Ecma-376.

## 5. What about PDF?

Portable Document Format (PDF) is a file format created by Adobe Systems in 1993 for document exchange. PDF is used for representing two-dimensional documents in a manner independent of the application software, hardware, and operating system. Based on Postscript Page Description Language (and thus not XML-based), PDF has been a *de facto* standard (with its specification known to public) for a long time). On July 1, 2008, PDF become an open standard by the ISO, as ISO 32000-1:2008.

Its popularity and accessibility on all platforms has made the PDF format a convenient publishing tool in situations where the end user does not need to edit the document.

PDF/A is a variant of the PDF format for the long-term archiving of electronic documents. PDF/A is an ISO standard, published in 2005, as ISO 19005-1:2005.

Finally, PDF/UA (PDF/Universal Accessibility) is a Standards Committee formed by AIIM. The mission of PDF/UA is to develop technical and other standards for the authoring, remediation and validation of PDF content to ensure accessibility for people that use assistive technology such as screen readers for users who are blind.

## 6. Conclusions and recommendations

In Section 2, I listed many examples of requirements and needs government agencies and public and private institutions look for in their search for open document formats. They all concentrate on the use of open standards, on cross-platform interoperability, vendor neutrality, and on being non binary (i.e., XML-based).

The recent approval (April 2008) of OOXML (ISO/IEC DIS 29500) resulted in existence of two XML-based office document standards (the other is ODF: ISO/IEC 26300, approved in May 2006) with overlap of 90%, and yet incompatible. The main reason for ECMA and ISO accepting OOXML for submission as an international standard was Microsoft's claim of its backwards compatibility with existing Microsoft proprietary binary formats. At the same time, OOXML approach, design and execution block full implementation by vendors/developers other than Microsoft. For further details, see the list of Technical disadvantages of OOXML, Google's and Oracle's comments in Section 4, article [22], very long list of problems listed in [21], and references in [23].

Furthermore, questions regarding Microsoft's Open Specification Promise (see [17]) raise uncertainty of the OOXML's legal status, and thus undermines its future implementations by

entities other than Microsoft. As noted in [24], and [25], Open Specification Promise (OSP) makes its promise "irrevocably," but only for *now*. Also, the OSP covers specifications, not a code, thus not permitting free software implementations. At this moment there are no OOXML implementations, even Microsoft Office 2007 does not support OOXML, as defined in ISO/IEC DIS 29500 or ECMA-376. These problems combined with OXML's complexity, extraordinary length (6000+ pages), technical omissions and single-vendor dependencies make alternative implementation unattractive as well as legally and practically impossible.

The other XML-based office document standard is ODF. Following [22],

- ODF is developed and maintained in an open, multi-vendor, multi-stakeholder process that protects against control by a single organization.
- ODF openness is reflected in the number of competing applications in which ODF is already implemented (see [13]).
- ODF is the only format unencumbered by intellectual property rights restrictions on its use in other software, as certified by the Software Freedom Law Center.
- ODF offers interoperability with ODF-compliant applications on most of the common operating system platforms.

Finally, in an unexpected move in May 2008, Microsoft announced (see [26]) that ODF (v1.1) will be a supported format in SP2 of Office 2007 (due first half of 2009). This is in addition to already supported PDF format in SP1 of Office 2007. Although it is not clear at this time what will be the quality/fidelity of this converter, Microsoft's move makes the ODF a clear choice for open document format.

At this time (November 2008), there are two doc-to-odf converters:

- Sun's plugin converter to/from odf (see, [27]);
- Microsoft's supported Open XML/ODF Add-in for Office (see, [28]).

Although the use of converters, in contrast to the use of native formats, should be depreciated, the above two converters provide a reasonably good transitional path to XML-based formats. A few comments about the above plugins. Sun's plugin works in all recent version of MS Office and is better integrated in MS Office menus; in addition, it allows for selecting ODF as a default format. On the other hand, Microsoft's supported plugin works only in MS Office 2007, but may provide better fidelity of conversion in some situations of the complex documents.

## Recommendations

(1) Use of the PDF format is recommended in situations where the end user does not need to edit the document.
(2) Use of the ODF format is recommended in situations where the edition of a document is needed.
(3) The OOXML format cannot be recommended for use at this time.

**Note:** *The above recommendations do* ***NOT*** *require purchasing of new software packages. Windows MS Office users can use the above plugins. Their installation is painless and their use doesn't require any habit changes. Mac's TextEdit exports/imports ODF since at least 2006, while Linux users never had problems with ODF. And besides, Open Office is freely available on all platforms for those who want to try something different for a change.*

## References

[1] ECMA-376: *Office Open File Formats* (OOXML): http://www.ecma-international.org/publications/standards/Ecma-376.htm

[2] ISO/IEC DIS 29500 *Office Open File Formats (OOXML)*:
http://www.iso.org/iso/pressrelease.htm?refid=Ref1123
[3] ISO/IEC 26300 *Open Document Format for Office Applications*:
http://www.iso.org/iso/pressrelease.htm?refid=Ref1004
[4] A Strategy for Openness. Enhancing E-records Access in New York state. Part I: Executive Summary:
http://www.oft.state.ny.us/Policy/ESRA/erecords/PartIerecordsStudy.pdf
[5] VALORIS, 2003. Comparative assessment of Open Documents Formats Market Overview. IA Brussels, Belgium: http://ec.europa.eu/idabc/en/document/3439/5585#VALORIS
[6] National Governments Requiring Use of ODF:
http://www.odfalliance.org/resources/Adoptions-ODF-Aug2008.pdf
[7] PEGSCO 2006, Conclusions and Recommendations on Open Document Formats, IDABC: Brussels, Belgium: http://ec.europa.eu/idabc/servlets/Doc?id=26971
[8] Government Open Source Policies, Center for Strategic and International Studies (August 2007):
http://www.csis.org/media/csis/pubs/070820_open_source_policies.pdf
[9] W. Ditch, XML-based Office document standards; in pdf format:
http://www.jisc.ac.uk/media/documents/techwatch/tsw0702pdf.pdf, or in ODF format,
http://www.jisc.ac.uk/media/documents/techwatch/tsw0702odt.odt.
[10] ODF Technical Committee home page:
http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office
[11] Open by Design:The Advantages of the OpenDocument Format (ODF). OASIS ODF Adoption TC, 10th December 2006:
http://www.oasis-open.org/committees/download.php/21450/oasis_odf_advantages_10dec2006.pdf
[12] OpenDocument v1.0 (Second Edition) specification. OASIS ODF Adoption TC, 19th July 2006:
http://www.oasis-open.org/committees/download.php/19274/OpenDocument-v1.0ed2-cs1.pdf
[13] Applications support for ODF:
http://www.odfalliance.org/resources/AppSupport20Dec2007.pdf
[14] OASIS, About Open Formula:
http://wiki.oasis-open.org/office/About_OpenFormula
[15] ODF Annual Report 2007:
http://www.odfalliance.org/resources/AnnualReport2007.pdf
[16] Google's Position on OOXML as a Proposed ISO Standard:
http://www.csun.edu/ hcmth008/odf/google_ooxml.pdf
[17] Microsoft Open Specification Promise:
http://www.microsoft.com/interop/osp/default.mspx
[18] Microsoft Office File Formats:
http://msdn.microsoft.com/en-us/library/cc313118.aspx
[19] Oracle: Unresolved Technical Concerns In DIS 29500 (OOXML):
http://www.odfalliance.org/resources/Oracle%20Technical%20Concerns%20DIS29500.pdf
[20] K. Ahrens, What about SVG? GullFOSS (Sun Microsystems blog, 2007):
http://blogs.sun.com/GullFOSS/entry/what_about_svg
[21] Grokdoc, EOOXML objections:
http://www.grokdoc.net/index.php/EOOXML_objections
[22] Sam Hiser, Achieving Openness: a closer look at ODF & OOXML:
http://www.odfalliance.org/resources/Achieving_Openness%20w-banner.pdf
[23] OOXML Analysis, ODF Alliance:
http://www.odfalliance.org/ooxml.php
[24] Microsoft's Open Specification Promise: No Assurance for GPL (Software Freedom Law Center):
http://www.odfalliance.org/resources/osp-gpl.pdf
[25] Interoperability woes with MS-OOXML (FSFE):
http://www.odfalliance.org/resources/msooxml-interoperability.pdf
[26] Microsoft Expands List of Formats Supported in Microsoft Office:
http://www.microsoft.com/Presspass/press/2008/may08/05-21ExpandedFormatsPR.mspx
[27] Sun's plugin converter to/from odf:
http://www.sun.com/software/star/odf_plugin/get.jsp
[28] Open XML/ODF Translator Add-ins for Office:
http://odf-converter.sourceforge.net

Mathematics, CSUN

*E-mail address*: jacek.polewczak@csun.edu