A Comparative Study for Open Document Formats

R. Mohamed Abdel Azim*, H. Farouk Ali**

Project Manager, eContent program

eContent Initiative Director, eContent program

Ministry of Communication & Information Technology (MCIT), Cairo, Egypt

Abstract — Many technical specifications that are sometimes considered standards are proprietary rather than being open, and are only available under restrictive contract terms (if they can be obtained at all) from the organization that owns the copyright for the specification. In this paper, we present a comparative study for the most common open document format, OASIS Open Document Format and Microsoft OXML

Index Terms **— Open Standards, XML, Open XML, Open Format, OASIS, Open Office**

I. INTRODUCTION

Open Standards are publicly available and implementable standards. By allowing anyone to obtain and implement the standard, they can increase compatibility between various hardware and software components, since anyone with the necessary technical know-how and resources can build products that work together with those of the other vendors that base their designs on the standard (although patent holders may impose "reasonable and non-discriminatory" royalty fees and other licensing terms on implementers of the standard).

Many technical specifications that are sometimes considered standards are proprietary rather than being open, and are only available under restrictive contract terms (if they can be obtained at all) from the organization that owns the copyright for the specification.

Being an open standard also does not necessarily imply that no licenses to patent rights are needed to use the standard or that such licenses are available for free. For example, the standards published by the major internationally recognized standards bodies such as the ITU, ISO, and IEC are ordinarily considered open, but may require patent licensing fees for implementation.

Open standards which can be implemented by anyone, without royalties or other restrictions, are sometimes referred to as open formats.

There is little really universal agreement about the usage of either of the terms "open" or "standard". Some people restrict their use of the term "open" to royalty-free technologies, while others do not; and some people restrict their use of the term "standard" to technologies approved by formalized committees that are open to participation by all interested parties and operate on a consensus basis, while others do not.

An open format is a published specification for storing digital data, usually maintained by a non-proprietary standards organization, and free of legal restrictions on use. For example, an open format must be implementable by both proprietary and free/open source software, using the typical licenses used by each. In contrast to open formats, proprietary formats are controlled and defined by private interests. Open formats are a subset of open standards.

The primary goal of open formats is to guarantee long-term access to data without current or future uncertainty with regard to legal rights or technical specification. A common secondary goal of open formats is to enable competition, instead of allowing a vendor's control over a proprietary format to inhibit use of competing products. Governments have increasingly shown an interest in open format issues.

In the context of business information exchanges, standardization refers to the process of developing data exchange standards for specific business processes using specific syntaxes. These standards are usually developed in voluntary consensus standards bodies such as the United Nations Center for Trade Facilitation and Electronic Business (UN/CEFACT)and the Organization for the Advancement of Structured Information Standards (OASIS).

II. *What is OpenDocument?*

Have you ever had trouble opening a document that someone sent you? Have you ever bought a copy of MS Office that you didn't want because you have to read documents that only work with MS Office? Have you ever wondered why there is so little choice in office software?

What you are seeing is vendor lock-in. It happens because your documents are written in a secret format that only one software maker knows. This prevents competitors from making products that can read and

write those files well. In short, it reduces your choices down to one.

Vendor lock-in is the enemy of competition. It short-circuits the market forces that would normally give you better products at a lower cost. OpenDocument is a way out of vendor lock-in for office software.

What if you could send a file to anyone and know that they can read it?

What if you could buy any product you want and know that you can still communicate with your customers?

This is the promise of the OpenDocument format. So, what is OpenDocument? OpenDocument is...

An open format. Any software maker can learn its details and make an application that can read and write this format.

An ISO standard. It is not controlled by one company, but by a non-profit standards group without a vested interest.

OpenDocument covers the features required by text, spreadsheets, charts, graphical documents and more.

III. PROS OF GOING "OPEN STANDARD"

First of all, organizations that store their data in an open format can avoid in this way being locked in to a single software vendor.

Second, you the user have finally the opportunity to end the long frustrating years in which Word or PowerPoint documents created on one PC would not be opened correctly by another PC with a different version of the program.

IV. TYPES OF OPEN DOCUMENT FORMAT

There are two competing XML-based formats for documents intended for use in office productivity software. These are OpenDocument and Microsoft Office Open XML. Both formats combine XML content with other media files into compressed archives (JAR in the case of OpenDocument, ZIP in the case of Office Open XML). In both formats, the main office document content and presentation information is stored as XML, with the ability to reference binary content such as BMP, GIF, JPEG. Both support the Dublin Core metadata standard.

There is fierce debate about technical merit between supporters of each format. A significant issue in terms of the success of the formats is the politics of adoption. The technical arguments, as in other battles for standards, often turn out to be less important than customer perception. Fundamental differences between the two formats are that OpenDocument is an approved ISO standard (approved for release as an ISO and IEC International Standard in May 2006, designated, ISO/IEC 26300) and is controlled by OASIS, a foundation broadly made up from representatives of the ICT industry and its customers. Microsoft Office Open XML is defined by Microsoft and currently undergoing a standardization process by Ecma International, an ICT industry standardizations organization. This Ecma standard will then be put through the process to gain ISO status. In the event of successful ISO adoption, control of the standard will then rest with Ecma International.

The OpenDocument format is implemented in several applications; at the time of writing Microsoft Office Open XML is being tested with beta versions of Microsoft Office 2007 and is a standard still in flux.

The OpenDocument format is the native format of both OpenOffice.org 2.0 and KDE KOffice, and is targeted as a native format for multiple applications. Microsoft Office Open XML will be used as the native format for Microsoft Office 2007. As well as Office 2007 providing native support for the format, a compatible plug-in will be released for some earlier editions of the suite. Those versions of Office will also receive a plug-in for OpenDocument support. It is not clear at this stage what level of interoperability either plug-ins will provide.[1]

V. OPENDOCUMENT OR ODF

About OASIS, OASIS (Organization for the Advancement of Structured Information Standards) is a not-for-profit, international consortium that drives the development, convergence, and adoption of e-business standards. Members themselves set the OASIS technical agenda, using a lightweight, open process expressly designed to promote industry consensus and unite disparate efforts. The consortium produces open standards for Web services, security, e-business, and standardization efforts in the public sector and for application-specific markets. Founded in 1993, OASIS has more than 4,000 participants representing over 600 organizations and individual members in 100 countries. Approved OASIS Standards include AVDL, CAP, DocBook, DSML, ebXML CPPA, ebXML Messaging, ebXML Registry, OpenDocument, SAML, SPML, UBL, UDDI, WSDM, WS-Reliability, WSRP, WS-Security, XACML, and XCBF. http://www.oasis-open.org/.

OpenDocument or ODF, short for the OASIS Open Document Format for Office Applications, is an open format for saving and exchanging office documents such as memos, reports, books, spreadsheets, databases, charts, and presentations. This standard was developed by the OASIS industry consortium and based upon the XML format originally created by OpenOffice.org. ODF was approved as an OASIS standard on May 1, 2005, and was approved for release as an ISO and IEC International Standard (ISO/IEC 26300) on May 8, 2006.

IBM, Sun Microsystems, and Others Develop Royalty-Free Standard for Office Applications Document Format and is publicly accessible. This means it can be implemented into any solution, be it open source or a closed proprietary product, without royalties. The OpenDocument format is intended to provide an open alternative to proprietary document formats so organizations and individuals can avoid being locked in to a single vendor.

OpenDocument provides a single XML schema for text, spreadsheets, charts, and graphical documents. It makes use of existing standards, such as HTML, SVG, XSL, SMIL, XLink, XForms, MathML, and the Dublin Core, wherever possible. OpenDocument has been designed as a package concept, enabling it to be used as a default file format for office applications with no increase in file size or loss of data integrity.

ODF is the first standard for editable office documents that has been vetted by an independent recognized standardization body.

Specifications, The most common file extensions used for OpenDocument documents are:

.odt for word processing (text) documents

.ods for spreadsheets

.odp for presentations

.odg for graphics

.odf for formulas (not a part of current Opendocument standard but a future extension)

An OpenDocument file can be either a simple XML file that uses <office:document> as the root element, or a ZIP compressed archive containing a number of files and directories. The ZIP-based format is used almost exclusively, since it can contain binary content and tends to be significantly smaller.

Standardization, The OpenDocument standard was developed by the OASIS industry consortium. The standardization process involved the developers of many office suites or related document systems. The first official OASIS meeting to discuss the standard was December 16, 2002; OASIS approved OpenDocument as an OASIS standard on May 1, 2005. OASIS submitted the ODF specification to ISO/IEC JTC1 on November 16, 2005, under Publicly Available Specification (PAS) rules.

After a six-month review period, on May 3, 2006 OpenDocument unanimously passed their six-month FDIS ballot in JTC1, with broad participation, after which the OpenDocument specification was "approved for release as an ISO and IEC International Standard" under the name ISO/IEC 26300.

After responding to all written ballot comments, and a 30-day default ballot, the OpenDocumement International Standard will go on to publication in ISO.

Further standardization work with OpenDocument includes:

OpenDocument 1.0 (second edition) has the status of a Committee Specification in OASIS. It includes all the editorial changes made to address JTC1 ballot comments

OpenDocument 1.1 is currently in a 60 day public review period in OASIS. It includes additional features to address accessibility concerns. OpenDocument 1.1 is expected by November 2006.

OpenDocument 1.2 is currently being written by the ODF TC. It will include additional accessibility features, metadata enhancements, spreadsheet formula specification based on the OpenFormula work (ODF 1.0 did not specify spreadsheet formulas in detail, leaving many aspects implementation-defined) as well as any errata submitted by the public. Originally OpenDocument 1.2 was expected by October 2007. However, upon learning that many of its activities will be completed far before then (e.g., the formula subcommittee expects to complete in December 2006), the group has agreed to develop a newer accelerated schedule.

Support for OpenDocument OASIS Standard, "IBM recognizes the importance of a standards-based document format. Use of open, non-proprietary formats will facilitate seamless collaboration between vendors, customers and partners and ensure the maintenance of corporate and government knowledge," said Karla Norsworthy, vice president, Software Standards, IBM. "IBM supports the OASIS OpenDocument formats, and we believe the standardization is a major accomplishment in an important area."

"Sun believes in the power of open standards to enhance business productivity and to stimulate innovation by preserving the intellectual property rights of content creators," said Tim Bray, Technology Director at Sun Microsystems. "Sun is a founding member of the OASIS OpenDocument Technical Committee, and Sun's StarOffice 8 productivity suite, based on the OpenOffice.org project, uses OpenDocument as its default file format.[2]

VI. OPEN XML (OOXML)

Microsoft Office Open XML (OOXML) is a file format developed by Microsoft to be used by the upcoming release of Microsoft Office 2007.

Microsoft's Office Open XML format uses a ZIP container for packaging XML and other data files. The resulting files are smaller than the binary files created by the previous Office formats. Microsoft maintains that its primary goal has to be backward compatibility with existing documents and full support of its extensive feature set. The Microsoft Office Open XML format is Microsoft's direct answer to the OpenDocument format (ISO/IEC DIS 26300) which was created by the OASIS foundation and uses similar technologies (XML

contained in ZIP). A comparison can be found in Comparison of OpenDocument and Microsoft XML formats.

File format and structure, The Open XML files consist of a ZIP package in which a set of individual XML files are placed that together form the basis of the Office document. Also included in the ZIP package will be embedded (binary) files like PNG, JPEG OR GIF images. A basic Open XML file contains an XML file called [Content_Types].xml at the root level of the ZIP package, along with three folders: _rels, docProps, and a directory specific for the document type (i.e. in a .docx word processing file that would be a word directory). The word directory will contain the basic

wordDocument.xml file which is the basis for the Office document. The directory in basic document wil vary depending on the type of office file created.

[Content_Types].xml file This file describes the content of the ZIP package. It also it contains a mapping for file extensions and overrides for specific URIs.

_rels Folder The _rels folders are where one goes to find the relationships for any given part within the package. To find the relationships for a specific part, one looks for the _rels folder that is a sibling of one's part. If the part has relationships, the _rels folder will contain a file that has one's original part name with a .rels appended to it. For example, if the content types part had any relationships, there would be a file called [Content_Types.xml.rels] inside the _rels folder.

Standardization, Microsoft has stated it will be an open standard, and has submitted it to the Ecma standardization process. The charter of the Ecma Technical Committee requires it to submit the completed standard to the ISO. Ecma announced on December 9, 2005 that it had accepted Microsoft's proposal to document the format as a proposed standard. It will be referred to as Ecma Office Open XML.

The Ecma technical committee developing the proposal includes representatives from Apple, the British Library, Canon, Intel, Microsoft, NextPage, Novell, Pioneer, Statoil ASA, Toshiba and The United States Library of Congress.

Since August 2006 Ecma is working on draft version 1.4 of the proposed standard. Also a liaison from the ISO/IEC from SC34 has been appointed to help prepare Open XML submission to ISO/IEC.

Licensing, The Microsoft Office Open XML format will be available under a free and perpetual license from Microsoft.

There has been a lot of argument about the ability for OSS software to use the format even under this fairly open license. Microsoft has tried to diminish these concerns by officially stating in a covenant not to sue that it will not sue any organization for using the format if the implementation complies to the official OOXML file formats. This has led to a greater reassurance that the OOXML formats will also be available for use in OSS software as even expressed by OSS licensing expert Larry Rosen.

A further indication of the totally free and open use of the format was given by Microsoft XML program manager Brian Jones as he presents a legal analysis on the convenient not to sue and also states that there is "no license needed to use the Office Open XML formats".

VII. ISO AND IEC APPROVE OPENDOCUMENT OASIS STANDARD FOR DATA INTEROPERABILITY OF OFFICE APPLICATIONS

The OpenDocument Format OASIS standard that enables users of varying office suites to exchange documents freely with one another has just been approved for release as an ISO and IEC International Standard.

OpenDocument, submitted by OASIS (Organization for the Advancement of Structured Information Standards), was balloted as an International Standard in ISO/IEC's Joint Technical Committee 1 on Information Technology. The standard has been given the designation, ISO/IEC 26300.

Most of today's electronic office documents have been created by a few commercial software programmes and more often than not each one has its own format. In order to process a document, users need the same program (and corresponding versions) or a filter that allows the document to be opened and modified. OpenDocument Format does away with this need.

The newly approved ISO/IEC 26300**,** *Open Document Format for Office Applications (OpenDocument) v1.0,* has been designed to be used as a default file format for office applications with no increase in file size or loss of data integrity. It will allow users to save and exchange editable office documents such as text documents (including memos, reports, and books), spreadsheets, databases, charts, and presentations – regardless of application or platform in which the files were created.

Organizations and individuals that store their data in the open format avoid being locked in to a single software vendor, leaving them free to switch software if their current vendor goes out-of-business, raises its prices, changes its software, or alters its licensing terms.

Billions of existing office documents will be able to be converted to the XML standard format with no loss of data, formatting, properties, or capabilities. This will facilitate document contents access, search, use, integration and development in new and innovative ways. [3]

"ISO/IEC 26300 is a shining example of what partnership in standardization can achieve for the

business community. Its publication underscores the importance of partnership among ISO and IEC and standards developing organizations such as OASIS to craft a common set of standards, and reflects the international community's recognition of the importance of open formats in enabling business interoperability," said Alan Bryden, ISO Secretary-General.

"ISO/IEC JTC 1's approval of OpenDocument as an International Standard is a major step forward in advancing the adoption of a format that gives all of us the flexibility to select the office application – commercial or open source – that best meets our needs," noted Patrick Gannon, president and CEO of OASIS. "We are particularly gratified by the broad range of national ballots cast in favor of the standard. This action underscores the international support for the OASIS open standards process that produced OpenDocument and delivers an assurance of long-term viability that is particularly important to governments."

ISO/IEC 26300 is the responsibility of ISO/IEC JTC 1, *Information technology*, subcommittee SC 34, *Document description and processing languages*. The standard will continue to be maintained and advanced by the OASIS OpenDocument Technical Committee and the recently formed OASIS ODF Adoption Committee, both of which remain open to participation from users, suppliers, government agencies, and individuals. [4]

VIII. COMPARING XML OFFICE DOCUMENT FORMATS.

OpenDocument	Microsoft OXML
Is an ISO standard	Is not a standard at all It does not have Ecma approval and has not even been submitted to ISO. Indeed Gartner predicts that ISO will not approve it.
Is vendor neutral	Is a one-company format The purpose of Ecma TC45 is to produce a format "that is fully compatible with [the format] submitted by Microsoft". In other words, they cannot make any substantive changes to the format. They can only rubber stamp it.
Many implementations applications list	**ZERO implementations** There isn't a single product in the market that implements the format. Not even from Microsoft.
5 years of development (in a standards body)	1 year of development (in a standards body)
Legible Readily intuitive to those familiar with HTML or DocBook.	Obscure See the technical comparison for details. The cryptic nature of OXML markup leads to higher development costs.
Proven technology Reuses proven standards like SVG and XLink.	Un-proven Reinvents the wheel.
Easier to implement 700 pages 3MB spec Reuses existing standards.	Harder to implement 4,000 pages 24.4MB spec Reinvents the wheel.

IX. COMPARING XML OFFICE DOCUMENT FORMATS: USING XML METRICS

First, a few words of caution. First, neither ODF nor MSOOX are completely finished or stable; the numbers may be different in 2008. Second, this is only one document from one provenance; the numbers may be different with the documents are entered native or come from different sources. Third, the files are the products of software, so to some extent they test the applications rather than formats per se; the numbers may be different for different applications. Fourth, the version of Word being used is a beta and some parts of Open Office are also probably immature: DOCBOOK export failed for example. (So to some extent this is a test of how some beta software produces data in a beta format, done to beta-test a utility using some beta metrics.[5]

Application Characteristics

Opened in Open Office, the document is about 736 pages. In Office 2007, the document formats at 732 pages. It doesn't seem a significant difference.

For load times, I logged off and logged on again to ensure a fresh session. I opened the applications and used the open menu rather than double clicking, so that application load time was not involved. For Open Office, the .SXW and .ODT files took about six seconds to load each (this was quite load dependent: at another time I noticed the same document taking about 14 seconds to load; I believe this may be due to Open Office being paged back into memory). For the Word 2007 beta, the (resaved) .DOC and .DOCX returned their initial page display faster than that: however the rest of the file loaded in the background and loading took about 35 and 45 seconds respectively.

Comment It seems that consideration of file loading needs to be slightly more nuanced than simple times

then: if you count to when you first see some text, Microsoft was much faster; however, if you count from when the document is fully loaded, Microsoft was significantly slower.

File Size

Here are the file sizes:
.SWX (original):434K
.ODT (ODF 1.0): 438K
content.xml (ODF): 4,383K
.DOC (MS): 4,432K
.DOCX (MSOOX): 764K
word/document.xml (MSOOX): 7,775K
.DOC (MS resave): 3, 142K
.DOCX (MSOOX resave): 733K
word/document.xml (MSOOX resave): 7,472K

Comment For a large files, the .ODT file is much smaller than the equivalent .DOCX file. This can be almost entirely attributed to the relative sizes of the XML files: the ODF XML file is much smaller than the equivalent MSOOX XML file. However, the differences in those files sizes are dwarfed by the difference in their size compared to the .DOC size which is five to ten times larger. Resaving the .DOC file resulted in approximately a 25% file size reduction.

XML Metrics

So here are the XML metrics.

Element and Attribute Count

Category	ODF	MSOOX (resave)
Element	103	95
Attributes	325	150
Total Metrics Value	428	245

Comments For the same document, MSOOX and ODF seem to require about the same number of unique elements. However, MSOOX has substantially fewer attributes required. (I will look further sometime, but I'd suspect that MSOOX is using richer data values rather than markup. It also seems that the ODF content.xml file contains more style information; both the ODF and MOOX ZIP structures have other files for containing style sheets.) At a minimum, we can say that processing ODF and MSOOX will involve different tasks: they are organized differently, and if the extra attributes in ODF are indeed due to a finer grain of markup then we can say that some kinds of document processing using XML APIs will be easier using ODF.

Field Count Metric

The field count metric here is a version of the field count metric presented in the blog before. The original metric required knowledge about which attributes were IDs, xml:space or other metadata, which requires a schema, annotations and perhaps some hand-counting. The metric here shortcuts matters by saying that the first attribute in each element is not a field.

Category	ODF	MSOOX (resave)
Number of Elements with Data Content (excluding Whitespace)	44213	90743
Number of Attributes (excluding First Attribute and Namespace Declarations)	12033	25407
Total Metrics Value	57246	121543

Comments The MSOOX numbers are about double those of the ODF. The reason for this seems to be that MSOOX uses an element value rather than attribute value for style information and something mysterious Bin64 encoded data called "fldData" (field Data) which are used for almost every chunk of text. I included this metric because I was concerned that Microsoft's highly nested style might inflate its document complexity metric, based on tests with tiny documents, but it turns out not to be the case.

Document Complexity Metric

Category	ODF	MSOOX (resave)
Element	103	95
Required Attributes	157	95
Optional Attributes	168	55
Required Children	16	19
Optional Children	112	73
Required as First Child	26	23
Total Metrics Value	582	360

Comment, According to these numbers, the ODF document is more complicated than the MSOOX document. This in part reflects the use of generic elements rather than specific elements, and as mentioned it may reflect a tendency in MSOOX to do more using rich data values rather than explicit markup.

Weighted Document Complexity Metric

The Topology Complexity Detective allows you to weight various factors to reflect the experience in your organization, when deriving metrics for cost or time estimation in projects. The following weighting is one such set, based on a particular client's experience for a certain kind of task.

Category	Weight	ODF	MSOOX (resave)

Element	2	103	95
Required Attributes	2	157	95
Optional Attributes	1	168	55
Required Children	1	16	19
Optional Children	1	112	73
Required as First Child	1	26	23
Total Metrics Value	-	842	550

Comment According to these numbers, the ODF document is more complicated than the MSOOX document.

What do these numbers mean?

The numbers seem to support the interpretation that beta MSOOX may be quite a bit less complex than ODF 1.0 at this stage, at least in the sense of using fixed structures more, and simpler in these sense of using fewer elements and attributes. ODF is flatter and has smaller file size but seems to include more style headers than the MOOX does. The metrics indicate that the use of attributes may be significantly different between the two formats, for example for people looking at data conversion estimation. On the application level, Open Office loads the ODT file much faster than the Word 2007 beta loads the DOCX file.

Finally, the fact the we have two (and presumably more) word processors that can produce and consume XML for a decent sized book, is such a great step forwards from a decade ago. A medal to both teams! Boiled down, based on these numbers (and I need to double check my thinking here, and this is completely blue sky!) I'd wouldn't be surprised if MSOOX were easier to convert *from* (because of its regularity, scale and low complexity) while ODF were easier to convert *into* (because of its richness and flexibility), after the initial hurdle of converting anything to/from either of them was leapt.

Conclusion

Regarding ISO approval *Open Document Format for Office Applications (OpenDocument) v1.0*, has been designed to be used as a default file format for office applications with no increase in file size or loss of data integrity. It will allow users to save and exchange editable office documents such as text documents (including memos, reports, and books), spreadsheets, databases, charts, and presentations – regardless of application or platform in which the files were created.

ACKNOWLEDGEMENT

The authors wish to acknowledge the assistance and support of the Information and Infrastructure sector and Arabic eContent Initiative at MCIT .

REFERENCES

[1] *OpenDocument software - Wikipedia, the free encyclopedia.htm "http://en.wikipedia.org/wiki/OpenDocument_software"* http://en.wikipedia.org/wiki/Comparison_of_OpenDocument_and_Microsoft_XML_formats

[2] *"OASIS Members Collaborate to Advance Open XML Format for Office Applications: Arbortext, Boeing, Corel, Drake Certivo, Sun Microsystems, and Others Develop Open Office Standard at Global Consortium" OASIS, 20 Nov 2002* http://www.oasis-open.org

[3] *ISO, "ISO and IEC approve OpenDocument OASIS standard for data interoperability of office applications", ISO Press Releases, www.iso.org. Retrieved on 2006-08-24.*

[4] *Open Document Format Gets ISO Approval* http://www.odfalliance.org/press/AllianceRelease3May06.pdf

[5] Rick Jelliffe, *"Comparing XML office document formats: using XML Metrics", Friday August 18, 2006 4:56AM,* http://www.oreillynet.com/xml/blog/2006/08/comparing_xml_office_document_3.html