From the Office Document Format Battlefield

Jirka Kosek
XML consultant

The two most common XML-based formats for office application suites—the Open Document Format and Office Open XML—are now international standards. This article describes the roads that lead to the creation and adoption of these two similar but imperfectly compatible formats.

If you regularly use an office application suite, you might have noticed changes over the past few years in the default formats for saving documents. From the user perspective, this isn't all that important—at least until you have problems opening a document from another system—but those small changes to the file extensions of office documents actually represent substantial changes to the underlying document representations. These changes are the result of a big movement in the IT industry. Indeed, it's a movement that could have a very interesting impact on how we handle digital content.

Recent versions of office suites are using new XML-based formats. OpenOffice.org and GoogleDocs, for example, use the Open Document Format (ODF) and Microsoft Office 2007 uses Office Open XML (OOXML). Both formats have now been accepted as international standards; this article outlines the history of the process that has left us with two functionally similar, but not fully compatible document formats.

History Lessons

During the 1990s, proprietary binary formats were the norm with office applications, and the developers rarely provided documentation that would let third-party software developers create applications that could consume or produce documents stored in those binary formats. Of course, various office suites provided support for Microsoft Office document formats, which dominated the market, but that support was always somehow limited.

The Internet boom of the late '90s brought greater attention to open source, open protocols, and open data formats as they helped establish the Internet as the platform of choice for communicating and sharing information. Yet, no such open format existed for general office documents.

About that time, developers of StarOffice (which later served as the base for the OpenOffice.org suite of office applications) started work on a new XML-based format for their applications. Seeing opportunities in this product space, several companies in the office applications area began working together under the auspices of the Organization for the Advancement of Structured Information Standards (OASIS; www.oasis-open.org) to develop the ODF standard. OASIS used the new XML-based OpenOffice.org format as the basis of ODF. (See Table 1 for a brief history of the ODF effort.)

Basing an office document format on XML has several advantages over using proprietary binary file structures. First, XML-based

 Published by the IEEE Computer Society 1520-9202/08/$25.00 © 2008 IEEE

Table 1. Open Document Format (ODF) history.

Date	Major events
1999	StarDivision—maker of the StarOffice suite, later purchased by Sun and open-sourced as OpenOffice.org—starts work on new XML-based file format.
December 2002	A new OASIS Open Office technical committee begins work on creating an XML specification for office document formats, starting from the XML-based format used in StarOffice/OpenOffice.org.
December 2004	OASIS renames the specification from Open Office Specification to Open Document Format for Office Applications (OpenDocument).
May 2005	OASIS approves ODF 1.0 as a standard.
September 2005	OASIS submits ODF 1.0 to the International Organization for Standardization/International Electrotechnical Commission Joint Technical Committee 1 (ISO/IEC JTC1) for approval through the JTC1 Publicly Available Specification (PAS) process.
October 2005	OpenOffice.org project releases OpenOffice.org 2.0, which uses ODF as its primary storage format
May 2006	ISO approves ODF 1.0 as ISO/IEC 26300.
November 2006	ISO/IEC 26300 is published.
February 2007	OASIS approves ODF 1.1 as a standard.

Table 2. Office Open XML (OOXML) history.

Date	Major events
1998–2000	Microsoft uses XML to represent some information in MS Office documents.
March 2001	MS Office XP is released with the ability to save spreadsheets in an XML-based format.
April 2003	MS Office 2003 is released with the ability to save documents and spreadsheets in an XML-based format. Soon after, Microsoft releases product schemas and documentation for newly created formats.
November 2005	Microsoft submits OOXML to ECMA for standardization. Built on previous XML-based formats used in MS Office, OOXML also includes XML representation for presentations.
November 2006	MS Office 2007 is released using OOXML as the primary storage format and with support for previous proprietary formats.
December 2006	OOXML is approved as ECMA Standard 376.
December 2006	ECMA International submits OOXML to the International Organization for Standardization/ International Electrotechnical Commission Joint Technical Committee 1 (ISO/IEC JTC 1) for approval via the JTC1 fast-track process.
September 2007	ISO members vote to deny OOXML approval. As a submitter of the standard, ECMA undertakes efforts to fix issues identified in ISO members' 1,027 distinct technical comments.
April 2008	Based on responses to comments provided by ECMA and the outcome of a five-day ballot-resolution meeting in Geneva, national bodies adjusted their initial votes and approved OOXML as the ISO/IEC standard 29500.

formats are transparent and thus quite easy to support in various applications. They're also easy to process with existing XML tools. This greatly simplifies the work of third-party developers who need to produce, consume, or otherwise interface with office documents.

Despite the promise of such advantages, Microsoft has been somewhat slow in implementing full support for XML-based formats into MS Office. In 2004, the European Union asked the company to publish its XML-based formats through a standards body. That combination of factors led to the creation of the Office Open XML (OOXML) format, which ECMA International (the European association for standardizing information and communication systems) later approved and submitted to the International Organization for Standardization/International Electrotechnical Commission Joint Technical Committee 1 (ISO/IEC JTC1) for approval. (See Table 2 for a brief history of OOXML.)

ODF and OOXML Internals

ODF and OOXML share many concepts. Both use XML to express document text and other important information, such as formatting and style definitions. Large embedded objects, such as images, are simply referenced from, rather than stored in, the XML content. To store the XML and embedded images as single files, both formats are, in fact, zip archives that use the format

developed by PKware (www.pkware.com). One benefit of that approach is that the text parts are efficiently compressed, so that the same document stored in one of these XML-based formats is usually smaller than a corresponding document saved in an older, uncompressed binary format.

ODF was developed and released earlier, but it's still somewhat immature in several respects. The OASIS ODF technical committee has been adding missing features gradually, but the format lacks various "enterprise" features, including standardized support for spreadsheet formulas and digital signatures. OASIS has already published ODF 1.1 and is currently working on version 1.2, which will finally add those missing features. But it remains unclear when (or if) the newer versions will also be approved as ISO standards.

Yet, ODF's relative simplicity has its advantages in terms of ease of implementation. Although Apple iPhone and some other high-profile applications and devices support OOXML, a greater number of applications currently supports ODF. On the other hand, OOXML was designed to capture all information stored in older Microsoft binary formats. Some have criticized the format's resulting complexity, but without features for handling legacy documents, it would be impossible to reliably convert existing binary office documents into more open, XML-based formats.

OOXML also includes features from the most advanced office suite (MS Office), so it's unlikely that users would need additional features anytime soon or that new releases of the OOXML format would be required as often.

Recognizing Common Office Format

The easiest way to differentiate between various office formats is to look at their filename extensions. The following table summarizes the most common file extensions in use.

Application type	MS Office Binary Formats	Open Document Format	Office Open XML
text	.doc	.odt	.docx
spreadsheet	.xls	.ods	.xlsx
presentation	.ppt	.odp	.pptx

For text documents, rich text format (.rtf) is also quite popular. Many applications support RTF, but the files can be bulky, especially with documents that contain images and other graphical elements.

Overview of Standardization Organizations

Many different organizations create standards that affect information technologies. The following list describes the most important organizations that deal with XML based formats:

- The International Organization for Standardization (ISO; www.iso.org) includes 157 national standards institutes, and each country has equal voting rights. In addition to information technology, ISO develops standards in areas such as quality, safety, and environmental protection.
- The International Electrotechnical Commission (IEC; www.iec.ch) is an organization that provides a platform for industry and academia to work together on education and research in the information industry. ISO and IEC cooperate through Joint Technical Committee 1 (JTC1), which is responsible for all IT-related standards within ISO and IEC. JTC1 was responsible for the fast-tracking of OOXML.
- The Organization for the Advancement of Structured Information Standards (OASIS; www.oasis-open.org) is a consortium of organizations and individuals that develops various XML-based formats. Membership to OASIS is open, and all dues-paying members can participate in the standardization process, although only organizations rather than individual members can vote.
- ECMA International (ECMA; www.ecma-international.org) is an industry association that develops standards in the areas of information and communication technologies and consumer electronics. Any company can join ECMA and participate in the standardization process by paying the membership fee.
- The World Wide Web Consortium (W3C; www.w3.org) develops standards for Web technologies. Any organization can join the W3C and participate in the standardization process by paying the membership fees.

ISO/IEC Standardization of OOXML

Over the past year, significant effort has gone into promoting OOXML as an international standard. The discussion has often been heated and polarized—somewhat uncommon with something as boring as a file-format standardization project. So, what caused such high interest? Skeptics would say, "money and business," and unfortunately, that's not far from reality.

Providers of competing commercial and open-source products have long been working to challenge MS Office's dominant position among office applications, but recent events have shown how hard it is to compete against Microsoft in that market.

ODF was initially created to become an OASIS standard, and OASIS later sent it to ISO/IEC JTC1 for approval as an international standard via the JTC 1 Publicly Available Specification (PAS) process. The PAS process went smoothly because, at the time, the ODF format wasn't very well-known (it was supported only in applications with insignificant market share). Consequently, many ISO/IEC JTC1 member countries essentially ignored the submission rather than giving it a vigorous technical review. (See the "Overview of Standardization Organizations" sidebar.)

The standardization of a format such as ODF was, of course, a step in the right direction, but its relative immaturity, feature deficiencies, and lack of acceptance by the world's largest office application provider were immediately problematic. At that point, some of Microsoft's competitors "hijacked" ODF to serve their own business goals. Many government agencies—particularly in Europe, but also including some US state and local governments—prefer to use software that conforms to ISO standards. Microsoft competitors thus started lobbying for their products by simply pointing out that they used ISO-approved formats, whereas Microsoft applications used proprietary binary formats.

For Microsoft, this was a challenging situation. Switching to

Answering CIOs' Most Frequent Questions about ODF and OOXML

Is there any general rule about choosing the Open Document Format (ODF) or Office Open XML (OOXML) as the new file format used in my company?

Neither format was built on a *green field*. Both were strongly influenced by existing products and formats. ODF has its origins in the StarOffice and OpenOffice.org office suites, whereas OOXML's origins are in the Microsoft Office suite. Your decision should be based on the software that's currently installed in your organization.

But don't forget that office suites aren't the only applications that deal with office documents. In many companies, documents are stored in various content-management systems, enterprise resource planning (ERP) systems generate reports directly in office formats, and so on. Before moving to a new file format, you should carefully analyze your interoperability requirements and the total impact of this change to your overall IT infrastructure.

Are there any risks connected with upgrading to ODF or OOXML?

Right now isn't great time to switch formats. I suggest waiting at least six to 12 months to see how the market evolves.

Developers are rapidly working on ODF, but it's still missing several features that are widely used in corporate environments. OASIS should be defining much of the missing functionality in version 1.2, to be released sometime in 2008. OASIS intends to send ODF 1.2 to ISO/IEC for approval via the PAS process. If all goes well, ODF 1.2 could be ratified as an international standard by 2009 or 2010.

OOXML was recently approved as the international standard, but many format changes were introduced during the process and it's not yet clear when Microsoft will implement them in its application suite.

Is it possible to use both formats at the same time? Are the formats interoperable?

For communication with external partners, supporting both new formats is ideal (for receiving as well as sending), but for internal communications adhering to one format is generally preferable to avoid increased costs for user training and IT support.

Interoperability between the formats is quite good, but imperfect. In practice, this means that you can lose some information or formatting can change during conversion. Various conversion tools are available, but they are still maturing. If your company doesn't have a serious need to quickly upgrade to a new file format, postponing the change a bit would be wise. We should see further improvements in conversion tools in 2008, as well as the release of a new version of OpenOffice.org with support for OOXML. It will take some time before other applications, such as content-management systems, report generators, and so on, support the new formats.

ODF wasn't viable for MS Office because of the format's missing features (not to mention the fact that its design was strongly influenced by competitive products). To defend its position, Microsoft had to create another document format that could get ISO approval while allowing a smooth transition from its feature rich proprietary format. The company decided to follow European Union preferences by opening and standardizing its office formats and sending the OOXML specification to ECMA for approval. Once OOXML was approved as ECMA 376, ECMA sent it to ISO/IEC JTC1 for approval via its fast-track process in which it is known as draft international standard (DIS) 29500.

The fast-track process is designed for standards that have already been ratified by a country member or by another standards-development organization that has "category A" liaison status with ISO/IEC JTC1. During this process, JTC 1 members can send comments and vote about proposed standards during a six month period. If the proposed standard doesn't gain enough support during that period, the submitter can try to resolve the comments. Resolution of comments and other modification of the proposed standard are then discussed and agreed during ballot resolution meeting (BRM). After the BRM, member bodies can change their initial votes throughout a 30-day period.

IBM, Google, and several other open-source proponents and Microsoft competitors strongly objected to DIS29500's approval. Some distasteful things appear to have occurred in several countries during the formation of national positions, and the draft was denied in the September 2007 balloting, during which national bodies submitted 3,522 comments (there were many duplicate comments, so the net outcome was actually 1,027 distinct comments). In January 2008, ECMA provided responses and proposed solutions to the majority of comments. The committee participants considered and further modified the proposed responses during the ballot-resolution meeting held in Geneva in February. National bodies had until 29 March to change their votes from the September BRM, and at the beginning of April ISO/IEC announced that OOXML was approved as international standard ISO/IEC 29500.

In the coming year, this situation should settle down as ODF and OOXML each establish their user bases. Both formats received significant attention over the past year, and a lot of research has gone into identifying the similarities and differences, as well as into developing conversion tools.

Many argue that it would be much better to adopt a single document format, but business interests appear to have made that impossible for now. ISO/IEC JTC1 considered both ODF and OOXML before they were fully ready. If both had spent a few more years in development at OASIS and ECMA, ODF and OOXML could have both been improved by, for example, including missing features and improving naming conventions and aligning with other related standards, letting market forces decide which format was more vital and thus ready for ISO/IEC approval. We missed that opportunity, but let's hope that at least some parallel universe has learned from our mistakes.

Acknowledgments

The views expressed in this article do not necessarily represent the IEEE's official position.

Jirka Kosek *is a teacher at the University of Economics in Prague. He has been a freelance XML consultant for more than 10 years. Kosek is an active member in several standardization bodies, including OASIS (DocBook and RELAX NG technical committees), W3C (Extensible Stylesheet Language and Internationalization Tag Set working groups), and ISO/IEC JTC1/SC34 (Document Schema Definition Language and topic maps). He is author of several books about Web technologies. In his free time he contributes code to the DocBook XSL stylesheets open-source project. Contact him at jirka@kosek.cz.*