Teksti: Mark Lange
Kimmo Bergius

Open XML and the Evolution of Document File Formats

Mark Lange
Senior Policy Counsel,
Microsoft EMEA
(mlange@microsoft.com)

Kimmo Bergius
Chief Security Adviser,
Microsoft Oy
(kimmob@microsoft.com)

When people use productivity applications such as Microsoft Office, Corel Word Perfect, or OpenOffice, they create, edit and save their work in files that include the author's text as well as graphic, and/or audio input in a specific file format that defines the structure and layout of that data. There are many document file formats used in the world today and most productivity applications have the ability to create, edit and save files using multiple formats. In addition, since the IT revolution began, the concepts of "document management" and "data exchange" have evolved, and will continue to do so, as innovation and consumer demand drive more efficient, productive methods and improved formats.

This year, there is a high-pitched debate about two specific file formats called Open XML and Open Document Format (ODF), both of which are relatively new open standards that are beginning to be widely used. The debate sometimes refers to a "standards war," as if it would lead to a single victor. Of all the real and imagined conflicts in life, it is hard to see how this one should provoke real tension. Few conflicts have so many elements of coexistence already in place. In fact, these and other file format standards serve a variety of customer needs and already are being implemented side by side. In many scenarios where users want to transfer data from one format to another, "translators" are rapidly developing to enable more complete, more transparent interoperability than has ever existed in the history of office productivity applications.

This article will explain more about Open XML as an open standard file format, in the context of the past, present and future of electronic document management. We will also discuss the uses of Open XML, and the ways in which it co-exists with ODF and other file formats.

The Past

When Microsoft first published the specifications for XML file formats in 2003, and licensed them freely to competitors, many people were surprised. People had become accustomed to the idea of that Microsoft Office file formats were "closed." However, in several respects, the introduction of XML formats marked a significant evolution in the state of the art of document processing.

Microsoft first started developing Office over 20 years ago. From the beginning, "binary" file formats (such as .doc or .xls) offered users rich features at relatively fast execution speed in computer systems with limited memory. By optimizing for speed and richness, there was a trade off in terms of data interoperability with other programs. These limitations are not restricted to Microsoft. Corel and Lotus have had the same issues with binary formats in their productivity applications. Despite the limitations associated with binary files, many software companies managed to achieve and advertise varying degrees of interoperability between different application environments with such files.

At the same time, Office and other applications provided users the ability to save files in "lowest common denominator" file formats like plain text (ASCII) and the rich text format (RTF) if they did not care as much about speed and richness and were more interested in the ability to have their data viewed, saved and reused in other contexts. HTML later emerged as a standard that was optimized for presentation and layout of documents on the Web.

The Present

Today, XML is rapidly becoming the industry standard for sophisticated document management and data exchange. XML is an open standard that was developed in the World Wide Web Consortium (W3C). Jean Paoli, who is General Manager for Interoperability and XML Architecture at Microsoft, was a co-creator of the XML 1.0 standard. With XML, data within one file can be easily accessed for reuse in a variety of ways, creating opportunities not possible with binary formats. Microsoft and many others in the industry believe that XML offers exceptional potential for automating virtually any task that involves working with data from documents (3). XML is used for presentation formats like Open XML and ODF, as well as any number of custom formats used for specialized purposes.

Microsoft began introducing XML-based formats in Office 2000 and dramatically increased this capability in Office 2003. From the beginning, the goal with these new XML formats was to offer users a way to take advantage of the power of XML and to give them a path to easily migrate their documents from binary formats to these XML formats while preserving the rich features they had used and rich content they had created in the past. With this technology advance, Microsoft also addressed public and private sector user requirements for greater interoperability and openness by publishing the specifications to its Office 2003 file formats on the Internet and freely licensing them.

The interoperability gains with XML-based formats were evident from 2003 (4), but there remained a desire, particularly among governments, that these formats be made open standards. Representatives of the member states of the European Union issued a set of recommendations in 2004 that specifically requested that Microsoft consider submitting its new formats to a standards body. Governments asked that Open XML be standardized because they wanted to ensure control over their own data and documents, now and for generations, independent of any single vendor's products. They felt strongly that control over the evolution of the file formats should rest with a standards body, and that the formats should be available for any vendor to implement without any restrictions.

These and additional requirements have been met with Open XML.

Open XML standardisation

In 2005, Microsoft joined numerous interested organizations in submitting Open XML to the standards body Ecma International to be considered as an open standard. To document the Open XML format, Ecma formed a technical committee including information technology companies (Apple, Intel, Novell, Microsoft, NextPage, Toshiba), government institutions that archive documents (the British Library, the U.S. Library of Congress) and users of information technology (BP, Statoil, Barclays Capital, Essilor).

These organizations had a common goal to define an XML format for sophisticated present and future uses that is also "capable of faithfully representing the pre-existing corpus of word-processing documents, presentations, and spreadsheets that are encoded in binary formats." (5) This was an important design goal because it is a specific and common customer need. For some data management scenarios, one could imagine that faithful rendering of legacy documents might not be a high priority, but for many users this is a significant requirement, especially as they look to migrate and preserve large archives of documents to XML-based formats. The Open XML specification was also designed to enable implementation on multiple operating systems and in heterogeneous environments, involving many types of data processing applications in a changing market.

The original specification submitted to the technical committee was approximately 2,000 pages. During the course of the standardisation work, this documentation grew to over 6,000 pages as a result of the committee's requirement that it comprehensively detail all aspects of the format – to provide competitors and users exactly what they need to implement it in full or in part (6). These demands were not surprising, given that the technical committee included companies that directly compete with Microsoft, and included users with very thorough requirements for long-term document management.

Microsoft clarified that any Microsoft patent needed to implement any part of the specification was available to anyone without restrictions under the Open Specification Promise, ensuring that any developer, including open source developers, can implement Open XML in their own programs (7).

On December 8th, 2006, Ecma approved the adoption of Open XML as an international open standard, labeled Ecma-376. This standards body also agreed to submit Open XML as a standard for ratification by ISO/IEC JTC1. Some governments had encouraged Ecma to seek this additional recognition and oversight of the standard by ISO. Open XML is now before ISO/IEC JTC1 for ratification.

Usage of Open XML

Beyond the usage of Open XML in Microsoft's current Office 2007 application (and the ability of users of prior Office versions to use Open XML through compatibility packs), developers are using Open XML in their own solutions and vendors of competing word processing applications are implementing the capability to work with Open XML files. In addition, a recent IDC Survey of companies in the Nordic region showed strong interest in Open XML among users interested in interoperability (8).

Novell offers a translator that provides support for opening and saving Open XML documents in OpenOffice. This implementation is based on the open source ODF-Open XML translator available on Source Forge. Corel has announced they will implement both Open XML and ODF in the Word-

Perfect program, stating that "this format-neutral approach allows Corel to focus directly on addressing the needs of customers, whose adoption choices will determine which formats will become most relevant." Also, Sun Microsystems is working on an Open XML import filter for spreadsheets.

Smaller companies are using the Open XML standard for a variety of implementations, some of which are posted on the OpenXMLDeveloper.org web site (9). An independent software vendor named Mindjet Labs has built solutions based on Open XML to expand the possibilities of its innovative MindManager program.

Custom solutions using Open XML can provide advances in efficiency, as experienced by the legislature in the state of Florida. Using Open XML, the Florida House of representatives is better able to manage amendments to bills in the legislative process.

The Future

"Both the ODF and the OpenXML document format specifications are XML based, promising great opportunities to explore the information contained in documents via tools other than traditional office suites. Examples of such exploration include indexing of document collections, automatic extraction of metadata from documents, search-engines, extraction of specific information for re-use, etc.." (IDABC Recommendations, 2006).

The "New World of Documents" is only beginning, and we believe that XML formats will spark an explosion of innovation and investment which, in turn, will bring great benefits for customers in the years to come. The industry has not yet defined or even imagined all the potential ways to aggregate and restructure data in documents, or discovered all possible workflow efficiencies. While standardized XML formats will become common in productivity suites, increasing importance will be placed on custom XML schemas for specialized business documents and eGovernment scenarios.

Multiple standards?

Open XML and ODF are already open standards and are already being deployed. With the freely available ODF-Open XML translator that is downloaded on our laptops, we have saved this article in an ODF format using our Microsoft Office application. This is not a situation where only one file format standard can physically exist or users are locked in to a single option. It is simply not a zero-sum game.

That is not to say that every document will, at all times, look or act perfectly the same when using different formats. That will be the case whenever we are dealing with different applications, sometimes even when using the same file formats, because applications have different features and capabilities, or different quality implementations of a single format.

Such variations have existed and will continue to exist, as long as there are multiple customer objectives, a competitive market, and evolving technologies. ODF and Open XML were created with very different design goals and they are only

Mark Lange explaining.

two of many document format standards in use today, each of which has characteristics that are attractive to different users in different scenarios. In particular, ODF was not designed to faithfully represent existing documents in binary formats, which is a real user need. At the same time, users' interoperability needs between the available formats are real. For this reason a great deal of investment is being made to satisfy that the multiple user requirements, and a great deal has been made possible with the current state of the art.

Multiple, co-existing standards are not unusual in the IT industry (10). For example, digital image formats, such as CGM, JPEG, and PNG, each of which is an ISO standard, meet different needs in the marketplace. When ODF was under consideration as an ISO standard, Microsoft made no effort to question that process because it fully recognizes customers' interest in the standardization of XML document formats and the potential for coexistence of multiple formats.

Some of the objections raised against the standardisation of Open XML come from the same people who in the past vigorously claimed to promote open standards, or who complained about Microsoft's "closed" binary formats. We may be witnessing a new category of "open double standards" applied by interests with a commercial competitive agenda. In this case, however, in seeking to work with others to standardize Open XML, Microsoft has made the right move for the right reasons.

* * * *

History has shown that customers want to use many different kinds of file formats because they have a diversity of needs and interests, and they also benefit from the evolution of file formats. Diversity and competition are good for customers because they allow them to choose packages that contain features that meet their individual needs. Increasingly, customers are able to choose products that implement all the file formats that they need, and the evolution of the technology enables them to discover more and different ways to manage their data in documents.

(3) See "A Foundation for the New World of Documents" (and Rick Jeliffe's analysis of that memo)

(4) See http://ec.europa.eu/idabc/en/document/2592/5588 ("The publication of the OpenOffice.Org and WordML formats has greatly improved the potential for interoperability of document processing.")

(5) See "Open XML Overview" from Ecma International.

(6) See http://tirania.org/blog/archive/2007/Jan-30.html ("Considering that for years we, the open source community, have been trying to extract as much information about protocols and file formats from Microsoft, this is actually a good thing.")

(7) See reference by Lawrence Rosen: "I see Microsoft's introduction of the OSP as a good step by Microsoft to further enable collaboration between software vendors and the open source community… I'm pleased that this OSP is compatible with free and open source licenses." http://www.microsoft.com/interop/osp/default.mspx

(8) See full report at http://openxmldeveloper.org/archive/2006/11/27/IDC_Open_Document_Standards.aspx

(9) See also Brian Jones' Open XML Formats blog for more examples.

(10) For another recent example regarding questions about the logic or necessity of insistence on one standard in an evolving market, see http://www.eetimes.com/news/latest/showArticle.jhtml?articleID=198100185

Other links

EU set of recommendations:
http://ec.europa.eu/idabc/en/document/2592/5588
Open Specification Promise:
http://www.microsoft.com/interop/osp/default.mspx
Ecma-376:
http://www.ecma-international.org/publications/standards/Ecma-376.htm
IDC Survey:
http://www.idc.com/nordic/about/press20061124.jsp
Novell translator:
http://download.novell.com/SummaryFree.jsp?buildid=ESrjfdE4U58~
ODF-Open XML translator available on Source Forge
http://sourceforge.net/projects/odf-converter/
Corel announcement:
http://www.corel.com/servlet/Satellite/us/en/Content/1153321430604?pressId=1164741065876
Sun: Open XML import filter for spreadsheets:
http://blogs.sun.com/GullFOSS/entry/office_open_xml_import_filter
http://notes2self.net/archive/2006/09/11/OpenXML-and-ODF-_2D00_-it_2700_s-not-a-zero-sum-game_2E00_.aspx
XML compatibility packs:
http://www.microsoft.com/downloads/details.aspx?familyid=941b3470-3ae9-4aee-8f43-c6bb74cd1466&displaylang=en

Mindjet:Labs' solutions:
http://mindjetlabs.com/cs/tags/Open+XML/default.aspx
Example from Florida:
http://www.microsoft.com/casestudies/casestudy.aspx?casestudyid=200428
IDABC Recommendations, 2006:
http://ec.europa.eu/idabc/servlets/Doc?id=26971
Not Zero sum game:
Different features and capabilities
http://www.koffice.org/filters/1.6/kword/oowriter.php
Different quality implementations:
http://www.opendocumentfellowship.org/applications
Open double standards
http://ntouk.com/?view=plink&id=251
Explaining commercial competitive agenda:
http://www.microsoft.com/interop/letters/choice.mspx