



ODF Europe Action Group - Making the Business Case for Business Ethics and Open Document Standards in Governments, Education, Business and Consumer Life

ODF Europe Action Group White Paper

– *ECMA claims explored and non-compliant examples shown*

– *ISO26300 standard pursues unrestricted universal legacy file preservation in practice and by intent*

- *Virtualisation technologies provide key to simplified legacy preservation but proprietary licence concerns highlighted*

Preserving legacy files with ECMA Office Open XML (MSOOXML)

Chris Puttick
CIO

Oxford Archaeology
April 2007

Just A Few of our Members

.riess applications gmbh (Germany) | 1dok.org (Germany) | 2shape GmbH (Germany) | 3BView Ltd (United Kingdom) | 60AT6 (United States) | Abshoff, Agricola, Prager & Venn GbR - freaque.net (Germany) | Access Foundation (United States) | Adept Softwares Pvt. Ltd. (India) | Adlib Software (Canada) | Adullact (France) | Advanced Information and Communication Technology (Iran) | Aero Systems Corp. (United States) | AFUL (France) | agelis Energieberatung (Germany) | Alfresco (United Kingdom) | Alka France (France) | All Stars s.r.o. (Czech Republic) | All Stars Sp. z o.o. (Poland) | Alma Technology (Australia) | Altar sp. z o.o. (Poland) | American Library Association (United States) | Amphora Research Systems (United States) | Anaska Formation (France) | Angulo Sólido (Portugal) | APITUX (France) | APPS Global Pty Ltd (Australia) | APRIL (France) | Ark Linux (Switzerland) | ARKNUS (Mexico) | Ars Aperta (France) | Asia OSPA Forum (India) | Asociación Mexicana Empresarial de Software Libre (Mexico) | ASS2L (France) | Association Lune Rouge (France) | Associazione Culturale Revolutionary Mind (Italy) | ATR, Inc. (United States) | Atviras kodas Lietuvai (Lithuania) | Auton Rijnsburg BV (The Netherlands) | Avanquest UK ltd (United Kingdom) | Axiros GmbH (Germany)...

Visit us at <http://www.odfalliance.org/memberlist.php> for the Full Member List. *You'll be amazed who's signed up. You can too.*

About ODF: To enable data users everywhere to have greater control over and direct management of their own records, information and documents, the ODF Alliance seeks to promote and advance the use of the ISO accepted Open Document Format (ODF) as the primary document format for governments, education, business and consumers

As documents and services are increasingly transformed from paper to electronic form, there is a growing problem that governments and their constituents may not be able to access, retrieve and use critical records, information and documents in the future.

The ODF alliance works globally to educate policymakers, IT administrators and the public on the benefits and opportunities of the Open Document Format, to help ensure that government and other user information, records and documents are fully and natively accessible across platforms and applications, even as technologies change.

Preserving legacy files with ECMA Office Open XML

Since 2006 there has existed a vendor-independent, ISO approved, fully documented standard for office documents, [ISO 26300](#), otherwise known as [Open Document Format](#). This standard is becoming [widely used](#) and [supported](#). Now a second standard for office documents is being proposed, [ECMA Office Open XML format \(EOOXML\)](#). In promoting the need for a secondary ISO standard for office documents, several points have been raised by proponents of EOOXML. On consideration, none of these points have proven particularly persuasive; the argument put forward addressed in this paper is downright spurious.

“OpenXML was designed from the start to be capable of faithfully representing the pre-existing corpus of word-processing documents, presentations, and spreadsheets that are encoded in binary formats defined by Microsoft Corporation.”

The above is verbatim from the [ECMA's overview document](#) for the EOOXML specification, and is defined as one of EOOXML's top reasons for being. This theme has been revisited during the 2007 abortive attempt to push EOOXML through the ISO fast track process, latterly in the Microsoft [open letter to IBM](#):

“Open XML... reflects the rich set of capabilities in Office 2007... and was designed to be backwards compatible with billions of existing [Microsoft Office] documents.”

The principle ECMA argument, but is it achieved?

Chris Capossela, a product management VP at Microsoft:

“all the features and functions of Office can be represented in XML and all your older Office documents can be moved from their binary formats into XML with 100 percent compatibility.”

And, finally, from the ECMA EOOXML approval press release:

“The Open XML standard recognizes the benefit of backward compatibility preservation of the billions of documents that have already been created while enabling new future applications of document technology.”

This example shows how in depth prior knowledge on one unpublished proprietary application is required

So the argument appears to be this:

that to ensure preservation of a series of undocumented legacy binary formats, it is necessary to have this new format that specifies, precisely and in great detail, all aspects of those legacy applications insofar as that applies to the layout and content of documents et al produced by said legacy applications, such that all aspects of these legacy documents can be reproduced.

Let this premise stand for the moment. Does EOOXML achieve this? In a hurried reading of the 6000 page document that is the EOOXML specification, it might appear that it does. A very slightly less hurried reading, and some issues become apparent; the more you look at those issues, the more significant they become. Take this example:

MSOOXML Legacy File Save Alert message

“2.15.3.6 autoSpaceLikeWord95 (Emulate Word 95 Full-Width Character Spacing) This element specifies that applications shall emulate the behavior of a previously existing word processing application (Microsoft Word 95)...”

*Our Future Document Heritage is
Digital*

So actually, despite 6000 pages, to achieve full backward compatibility for a Microsoft Word document you have to already know how Microsoft Word 95 works. A thorough examination discovers other similar clauses, with the end result that to implement EOOXML so that it satisfies the full legacy format support promised for Microsoft Word documents, you need to not just know how Word 95 works, but also how generations of Microsoft Word work: Microsoft Word for Macintosh 5, Microsoft Word for Windows 6, 97, 2002, 2003, etc..

*MSOOXML is not all embracing by
design*

Let us try another test; if the intent of EOOXML is full compatibility with legacy files, and Microsoft Office 2007 stands as the sole implementation of the specification, then we should be able to open a legacy file and save it in the new format without loss of data. A possibly lengthy series of tests to fully explore if this holds true. Fortunately it takes only opening and saving a few files before we can show that it is not. One expert, Rob Weir, trying Excel 2007, opened a legacy Excel file, and on saving it in EOOXML received a message:

“This [Excel] document contains one or more of the following features that are not supported by the selected file format [EOOXML]”.

See Rob Weir's [blog entry](#) for more information.

So, EOOXML not only fails to provide sufficient information to provide 100% fidelity for legacy formats, but the reference implementation, by the only organisation that could and should know all the missing information, fails a simple test. In fact it gets worse; Directions on Microsoft, independent Microsoft analysts, point out that [Microsoft Word 2007 does not support](#)

[versioning](#) within legacy files. Whether or not EOOXML can reproduce that versioning, the reference implementation cannot; an important preservation feature, that may tell you much more about a document than even its final version content, is unsupported.

Now let us return to the premise that a new format was needed to ensure preservation of legacy files with 100% fidelity. First, let us all agree – there is a need for preservation of legacy digital material. Outside of regulatory requirements, which are many and growing, there is also a wealth of future heritage material that is now being born digital – the formats used are varied, and some of them might not come to light for many years to come. So is there a need for a new file format to ensure these legacy documents are preserved?

Let us look at the position we now find ourselves. There is a raft of documents out there, some of which are in formats that we just cannot read. These are a tiny minority which date from early in the use of computers to store files in anything more advanced than text, and the EOOXML specification does not deal with them. Then there are the office category files created as home and corporate computers became more common; these will be in varied formats and from various platforms – Apple Macs and AppleWorks, Amstrad PCWs and Locoscript, IBM PCs and Wordstar, etc.. Following that era are files from early versions of Microsoft Office and WordPerfect, AmiPro etc., and then we get to the last ten years, where the majority of files are in various Microsoft Office formats, or Corel Office, Lotus SmartSuite, OpenOffice etc.. The EOOXML specification is designed to be backwardly compatible with this latter category, and, by stated intent, only those developed by Microsoft.

Example:

Virtualisation Technology simplifies Information Preservation Virtualisation technology is a way forward to legacy preservation with true fidelity. By preserving data (document, spreadsheet, videoclip) along with its environment, (computer, application program). It requires vendors to grant licences in perpetuity to all comers, which could be easily achieved.

What this backward compatibility offers is, for example, the opportunity to save a MS Word 95 in ECMA OOXML, yet when reopened have it look like it did in Word 95. This might be useful for archives and libraries who wish to slavishly recreate the digital document as viewed by its creator, but is it much use to the general populace? It might, perhaps, be useful for a complex document with formatting that might transform meaning if migrated to another format, but consider that the EOOXML specification is over 6000 pages long, and still it doesn't contain all the information needed to implement this backward compatibility; that must then be in the Microsoft Office 2007 software.

So this brings us to another question: is it actually the application rather than the format that is most important? To be sure, to have the capacity to represent in a new format everything that was possible in previous ones is necessary for full fidelity, but this does not actually provide full fidelity. Indeed, one could have two essentially feature-identical formats and still not get 100% fidelity. Although it would be possible to create an application that read both and therefore could faithfully represent, with 100% fidelity, files created in a second application, it still requires that the application understands both. Moreover, for 100% fidelity, the new application must understand precisely how each feature present in the file was represented by the originator application.

Conclusions are clear

With all these challenges to overcome, is 100% fidelity presentation of an undocumented legacy format possible? Do we need the greatest gift of the faerie given to us to see a file as the author saw it? In terms of creation of a new application that can read that legacy file, it is not an impossible feat. Just, very, very nearly.

You need several ingredients:

- (a) one intimate working knowledge of the application that created the legacy file;
- (b) one copy of said application;
- (c) one copy of the author's computing environment (including screen resolution, printer drivers, fonts, etc.);
- (d) lots and lots of time.

Of course, out of those ingredients there is another easier cake to be baked. If you want 100% fidelity, to see the file as the author saw it, you need only ingredients (b) and (c). While for those formats which even the platform on which the application that created them is lost in the dawn of computing time these are not available, for those files created by Microsoft Office it is not only entirely possible but in fact rather simple to recreate the environment, utilising virtualisation technology. Preserving the original application along with the environment effectively guarantees 100% fidelity, in a way no other approach can; and it is simple, quick and easy to do with freely available tools. If 100% fidelity of legacy documents is what Microsoft are keen to ensure, they need merely to grant licences in perpetuity to all comers for use of these legacy applications and associated environments.

Another question to be addressed: is 100% fidelity necessary or even useful? For something like CAD, it would seem important. For word processed documents, spreadsheets, presentations? Maybe, but the need is not so clear nor widespread. Let us examine some different audiences interested in preservation of office documents.

Government agencies

Concerns include short- to medium-term access for business and compliance needs. Some proportion of documents retained indefinitely for archival purposes.

Businesses

Concerns include short- to medium-term access for business and compliance needs. Some proportion of documents may be retained indefinitely for archival purposes.

Third sector organisations

Concerns include short- to medium-term access for business and compliance needs. Some proportion of documents may be retained indefinitely for archival purposes. In the academic sector old research may form the basis of new research and so continuing accessibility is important.

Individuals

Concerns include short- to medium-term access for personal and professional needs. Some proportion of documents may be retained indefinitely and even passed on to family members.

For how many of the documents produced by these groups is exact reproduction necessary, rather than preservation of content and meaning? For that small number, would use of [PDF/A](#) (ISO 19005-1:2005) be more appropriate? This guarantees a visually identical document in a preservable format, and moreover one for which any number of free readers exist. In fact, in those cases, would a migrated document be acceptable? Surely only a digitally authenticated and author verified PDF/A document, created within the same time-space or a copy of the original application and environment would be deemed acceptable. For the remainder it is only content and meaning that is of interest. This brings us

to our penultimate question: Is ISO26300 capable of storing the content and meaning of converted legacy documents as originally intended?

In almost all cases, the answer is a clear and resounding yes. There are a limited number of exceptions all of which are a result of undocumented aspects of those previous specifications. These will be covered in a future (100% backwardly compatible!) generation of ISO 26300, currently under development at OASIS.

So in conclusion it seems that:

- there is almost no need for the preservation capabilities claimed for ECMA Office Open XML;
- the 100% legacy support claims are in fact untrue;
- planned developments of the existing ISO 26300 standard will shortly provide for the scarce few cases where documents in a legacy format cannot currently be translated into ISO 26300 format without loss of meaning, and finally;
- commonly available virtualisation technologies can provide a more effective and safer solution where true 100% fidelity is needed or desired.

So to close with the final question: tell me again, why is it that we need a secondary standard for office documents?

References

<http://www.ecma-international.org/publications/standards/Ecma-376.htm>

<http://www.ecma-international.org/news/TC45>

[current_work/OpenXML%20White%20Paper.pdf](http://www.ecma-international.org/news/TC45_current_work/OpenXML%20White%20Paper.pdf)

<http://www.microsoft.com/interoper/letters/choice.mspx>

<http://www.robweir.com/blog/2007/01/formats-of-excel-2007.html>

<http://www.directionsonmicrosoft.com/sample/DOMIS/update/2007/01jan/0107nffio2.htm>

Points raised by proponents of the ECMA Office Open XML format (MSOOXML).

ECMA's overview document for the MSOOXML specification

Microsoft open letter to IBM:

Opening a legacy file in Excel 2007, and on trying to save it in MSOOXML received a message "This document contains one or more of the following features that are not supported by the selected file format".

Directions on Microsoft, independent Microsoft analysts, pointed out that Microsoft Word 2007 does not support versioning within legacy files.