



MS Office Open XML

Flaws in the current standards proposal

Arnd Layer, IBM Germany
2007-06-19

Agenda



- Definition of standards and requirements emerging thereof
- How does OOXML fulfill those requirements?
- Summary

What is a Standard?

- “[A] document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context
NOTE Standards should be based on the consolidated results of science, technology and experience, and aimed at the promotion of optimum community benefits.”
 - ISO/IEC Guide 2:2004, definition 3.2

- “...a standard is an agreed, repeatable way of doing something. It is a published document that contains a technical specification or other precise criteria designed to be used consistently as a rule, guideline, or definition. Standards help to make life simpler and to increase the reliability and the effectiveness of many goods and services we use. They are intended to be aspirational - a summary of good and best practice rather than general practice. Standards are created by bringing together the experience and expertise of all interested parties such as the producers, sellers, buyers, users and regulators of a particular material, product, process or service.”
 - <http://www.bsi-global.com/en/Standards-and-Publications/About-standards/What-is-a-standard/>

- “A purpose of IT standardization is to ensure that products available in the marketplace have characteristics of interoperability, portability and cultural and linguistic adaptability. Therefore, standards which are developed shall reflect the requirements of the following Common Strategic Characteristics:
 - Interoperability;
 - Portability;
 - Cultural and linguistic adaptability.”
 - JTC1 Directives, 5th Edition, Version 3.0, Section 1.2

What is a Standard?

- “[A] document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines or characteristics for activities or their results, aimed at the achievement of the optimum degree of order in a given context
NOTE Standards should be based on the consolidated results of science, technology and experience, and aimed at the promotion of optimum community benefits.”
 - ISO/IEC Guide 2:2004, definition 3.2

What is a Standard?

- “...a standard is an agreed, repeatable way of doing something. It is a published document that contains a technical specification or other precise criteria designed to be used consistently as a rule, guideline, or definition. Standards help to make life simpler and to increase the reliability and the effectiveness of many goods and services we use. They are intended to be aspirational - a summary of good and best practice rather than general practice. Standards are created by bringing together the experience and expertise of all interested parties such as the producers, sellers, buyers, users and regulators of a particular material, product, process or service.”
- <http://www.bsi-global.com/en/Standards-and-Publications/About-standards/What-is-a-standard/>

What is a Standard?

- “A purpose of IT standardization is to ensure that products available in the marketplace have characteristics of interoperability, portability and cultural and linguistic adaptability. Therefore, standards which are developed shall reflect the requirements of the following Common Strategic Characteristics:
 - Interoperability;
 - Portability;
 - Cultural and linguistic adaptability.”

- JTC1 Directives, 5th Edition, Version 3.0, Section 1.2

Common Themes for Standards

From these and other national definitions, some common themes emerge on what qualities standards should have:

- 1.They define **precise common** criteria for doing something in a **repeatable** way
- 2.They provides optimal degree of order in a given context, intended to be **aspirational**, giving the **consolidated** results of science, technology and experience, a summary of good and **best practice** rather than general practice.
- 3.They encourage **interoperability** and **portability**
- 4.They have **cultural and linguistic adaptability**.

Precise, Repeatable, Common (I)

- The OOXML specification merely lists the names of these settings. It does not define them. Although Microsoft is the only one who knows what these settings mean, they decline to give a precise definition of these attributes. They are only defined, indirectly, by reference to legacy software applications:
- Note that the primary goal and benefit of OOXML is (according to its authors):
 - *“...to enable the implementation of the Office Open XML formats by the widest set of tools and platforms, fostering interoperability across office productivity applications and line-of-business systems, as well as to support and strengthen document archival and preservation, all in a way that is fully compatible with the large existing investments in Microsoft Office documents.”*
- “Compatibility Settings”:
 - “footnoteLayoutLikeWW8”
 - “autoSpaceLikeWord95”
 - “useWord97LineBreakRules”
- *“To faithfully replicate this behavior, applications must imitate the behavior of that application, which involves many possible behaviors and cannot be faithfully placed into narrative for this Office Open XML Standard. If applications wish to match this behavior, they must utilize and duplicate the output of those applications.”*
- **Being “fully compatible with the large existing investments in Microsoft Office documents”**
 - **Is reserved for Microsoft alone.**
 - **The OOXML standard does not provide for repeatable or common practice of this benefit.**

Precise, Repeatable, Common (II)

However impressive the list of numeration conventions listed, the fact is that

- they are mere labels, and
- are not precisely defined .

The reader of the OOXML specification is told that something called “Korean Legal Numbering” exists, but they are not told what it means or how to practice it in their application.

- **The OOXML specification simply does not provide for repeatable, common use of these features.**

- Numbered list styles representing in WordProcessingML

- “chicago”
 - *“that the sequence shall consist of characters as defined in the Chicago Manual of Style”*
 - without specifying an edition of that manual (there have been 15 editions of The Chicago Manual of Style) or a page reference.
- “ideographDigital”
- “ideographLegalTraditional”
- “koreanDigital2”
- “koreanLegal”.

Precise, Repeatable, Common (Summary)

These criteria speak to the need for a standard to provide a detailed, written description that allows for the common practice of the technology.

In summary, we found that many areas of OOXML are undefined. Although the specification does provide a formidable framework for Microsoft to represent its own documents in, this ability does not translate into anything approaching equal access for others to obtain these benefits.

The question for reviewers to ask is not, “Does OOXML represent the features of billions of legacy Office documents?”

The question to ask is, **“Does OOXML define its file format in a precise way that allows repeatable and common practice of its claimed benefits?”**

Examples

- Compatibility Settings
- List Styles
- Security Descriptor

When we examine OOXML under this light, we find that, in its present form, it fails to satisfy that criterion. Its lack of maturity as a standard, reflected also in the lack of multiple full-featured implementations, and insufficient prior technical review make it inappropriate for Fast Track consideration, and in its present state, unacceptable for approval as an International Standard.

Aspirational, Consolidated Best Practices

An ISO Standard should not merely be the minutely detailed record of the operating characteristics of a single company's product, no matter how dominant that company is in their field. An International Standard should represent the “consolidated results of science, technology and industry”. It should be “aspirational”. In other words, it should not just show one vendor's way of accomplishing a task. It should attempt to provide “a summary of good and best practice” based on the consensus of expert opinion. It should teach the best practices for the repeatable, common practice of a given technology.

One way our industry records its best practices is through standardization. So the existing body of document and markup standards represents a compendium of already reviewed, approved, and implemented best practices. The work of the World Wide Web Consortium (W3C)¹ is especially relevant to XML document formats, since they maintain the core XML standard as well as related standards such as XHTML, CSS2, XSL, XPath, XForms, SVG, MathML, SOAP, the standards that represent the very backbone of XML and XML-related technologies.

When we look at OOXML, however, we see very little reuse of the consolidated best practices of the industry. In fact, we find the opposite, that Microsoft is proposing the adoption of legacy formats used by them and only them, even when relevant W3C standards are at hand.

Aspirational, Consolidated Best Practices (I)

VML was developed by Microsoft and proposed by them to the W3C, where it was evaluated by a technical committee and rejected. This was in 1998. Instead the industry decided to back Scalable Vector Graphics (SVG) which was developed into a standard by the W3C and widely adopted. The standard for XML vector graphics has been SVG for almost a decade. The only reason why VML shows up in OOXML today is because Microsoft made the wrong bet, and integrated VML rather than SVG into Internet Explorer and Office 2000.

We should note that Microsoft is aware that VML is the wrong standard to use for vector graphics (quote).

So instead of using the existing standard SVG, Microsoft is including two different markup languages for vector graphics, one which was rejected in 1998 by the W3C, and one which they developed in isolation. The amount of extra work this causes for everyone who implements OOXML is immense. They will need to support two different markups for the same thing, even though this gives no additional benefit to their users. Well, everyone, but Microsoft, of course, since they already have legacy support for VML in their product.

VML

“The VML format is a legacy format originally introduced with Office 2000 and is included and fully defined in this Standard for backwards compatibility reasons. The DrawingML format is a newer and richer format created with the goal of eventually replacing any uses of VML in the Office Open XML formats. VML should be considered a deprecated format included in Office Open XML for legacy reasons only and new applications that need a file format for drawings are strongly encouraged to use preferentially DrawingML”

Does this sound aspirational?

Does this sound like they are promoting best practices?

On the contrary, they have added 600 pages of VML requirements to the OOXML specification that bring no value to anyone but Microsoft, and in fact creates steep barriers to others who would implement OOXML.

Aspirational, Consolidated Best Practices (II)

In other words, the Gregorian Calendar, the base calendar of commerce, science and government worldwide, is set aside, for “legacy reasons”. The result is that all implementors of OOXML required to have their applications give their users incorrect answers to questions like “What day of the week in February 1st, 1900?”, if they want to be conform to the OOXML standard. This causes particular pain in the common task of exchanging spreadsheet data with relational databases via SQL, a standard which explicitly requires the use of the Gregorian calendar.

Backwards compatibility could also be established by converting the dates when saving in OOXML. Why repeat a known fault in a new standard?

Date Format in Spreadsheets

“For legacy reasons, an implementation using the 1900 date base system shall treat 1900 as though it was a leap year... A consequence of this is that for dates between January 1 and February 28, WEEKDAY shall return a value for the day immediately prior to the correct day, so that the (non-existent) date February 29 has a day-of-the-week that immediately follows that of February 28, and immediately precedes that of March 1.”

Is this a best practice?

Is this the consolidated result of science, technology and experience?

Is this what we should aspire to?

Aspirational, Consolidated Best Practices (III)

OOXML defines a new string type called “*Basic String*,” “*a binary basic string variant type*.” One of the features of this new string type is that it allows non-XML characters (control characters) to be specially encoded. However, the presence of non-XML characters in an XML document, is contrary to interoperability of XML and XML-based tools.

The W3C’s Internationalization Activity:

“Control codes should be replaced with appropriate markup. Since XML provides a standard way of encoding structured data, representing control codes other than as markup would undo the actual advantages of using XML. Use of control codes in HTML and XHTML is never appropriate, since these markup languages are for representing text, not data.”

Aspirational, Consolidated Best Practices (IV)

In several places⁷ OOXML makes use of 'bitmasks' to encode multiple boolean values into a single integer. This is a practice common in constrained memory environments in programming languages like C. However, it is considered very bad style in XML, since it makes processing by standard XML tools like XSLT extremely difficult, since these tools lack bit-level operations needed to effectively process data at the bit level.

Not only does OOXML fail to provide a consolidation of best practices in a problem domain from science, industry and experience, it fails to even provide a consolidation of Microsoft's own best practices. For example, OOXML recommends that print settings (number of pages to print, which pages to print, orientation, print quality, etc.) be stored in a platform-specific binary format. For example on Windows their guidance is to store in what is called the "DEVMODE" structure.⁸ Doing so would render the print settings platform dependent and hurt interoperability. But at the same time, Microsoft's new specification, "XML Paper Specification" (XPS) offers a PrintTicket element of which Microsoft says:

MS XML Paper Specification (XPS):

"PrintTicket technology is the successor of the current DEVMODE structure. It is an eXtensible Markup Language based document that specifies and persists information about job formatting and print job configuration.... Relative to the current print subsystem, the PrintTicket technology enables all components and clients of the print subsystem to have transparent access to the information currently stored in the public and private portions of the DEVMODE structure, using a well-defined XML format."

Why is OOXML getting the inferior, binary, unportable, platform- and application-dependent print settings, when Microsoft's own best practice is to move to a "well-defined XML format"?

Aspirational, Consolidated Best Practices (V)

Similarly, OOXML defines several cryptographic algorithms which are not among those approved for use by NIST in their FIPS-180 list of compliant algorithms¹¹. Instead of using a recommended algorithm like SHA-256, Microsoft specifies a legacy hashing algorithm, presumably used in earlier versions of Microsoft Office.

Does this teach the consolidated best practices of science, industry and experience?

On the contrary, Microsoft doesn't even recommend using these algorithms.

Instead, they provide DRM-based protections in Office 2007. These are not documented in OOXML, so no other vendor is able to freely use those features. Instead we're left with the flawed legacy security support of OOXML, support which is not even FIPS-180 compliant.

So in this example, Microsoft is keeping the best practices to themselves, and leaving the OOXML specification with crippled security.

Aspirational, Consolidated Best Practices (Summary)

In summary our observation is that OOXML is a literal porting of the features of a single vendor's binary document formats. The avoidance of relevant existing international standards, as well as the inconsistent use of that vendor's own preferred technologies demonstrates that OOXML does not represent the consolidated results of science, industry and experience.

- Examples
 - VML
 - Date Format
 - Basic Text
 - Bitmasks
 - Cryptographic Algorithms

It is not aspirational.

Although it may provide a technique of reading data in that one vendor's format, that at best recommends it as only a technical specification.

But since it does it does not represent the consolidated best practices in the industry, a defining quality of an ISO standard, we cannot recommend approval of this ballot.

Interoperable & Portable

Portability and Interoperability are two of JTC1's "Common Strategic Characteristics" and as such are requirements of all JTC1-approved standards. In the realm of document format standards, questions for us to ask are whether the proposed document format can be fully implemented by multiple applications on multiple operating systems? Or has it been written exclusively for the benefit of a single vendor's application?

We have not found that OOXML satisfies the requirements for portability and interoperability. On the contrary, OOXML is heavily tied to the Microsoft Office applications and to Microsoft Windows, to the detriment of interoperability and portability.

Interoperable & Portable (I)

Data Interchange

between Spreadsheets and RDBMS

An important area of interoperability is the interchange of data between spreadsheets and relational databases. This is a common occurrence and has been supported by most spreadsheet vendors for over a decade. However, as defined, OOXML has no way to represent dates before the year 1900, while modern databases can represent much earlier years. IBM's DB2 can support dates to the year 1, for example. And Oracle supports dates back to the year 4712 B.C..

Although no one will deny Microsoft the right to ship a product with a lesser date restriction if they want, the OOXML standard should not prevent other vendors from doing a better job. An application vendor will naturally want to match their spreadsheet's date support to the equivalent capabilities of their database. So why is OOXML restricted to the limitations of Microsoft Excel? This hurts interoperability between spreadsheets and databases.

Clipboard Formats

OOXML defines a ST_CF type2, which records the allowed clipboard formats which may be used with a graphical object. The allowed values of this type, EMF, WMF, etc., are all proprietary Windows formats. No allowance seems to have been made for use by other operating systems.

For example, in Linux images are typically copied on the clipboard in an open standard format like PNG. But if a vendor encodes "PNG" into a document record of this type, the document will not be valid and the application therefore will no longer conform with the OOXML standard.

Interoperable & Portable (II)

Password Hashing

The definition of a password hashing algorithm in SpreadsheetML is given by presenting 5-pages of C-language source code³, likely extracted from Excel. However, the bit manipulations of this code is inherently machine dependent, and will give different results depending on the processor architecture.

It is not a portable definition of the function.

Optimize for Browser

The “optimizeForBrowser” element of WordProcessingML4 has been defined in a way which ignores the existence of current browsers other than Internet Explorer.

What about Firefox?

What about Safari?

What about Opera?

None of these can be set as target browsers.

This section requires that “*all settings which are not compatible with the target web browser shall be disabled.*” But what if I want my application to produce standards-compliant output? So yes to PNG, no to VML, yes to MathML and SVG?

A vendor is not able to specify this with the way OOXML has been designed.

Interoperable & Portable (III)

Slide Synchronization Properties

The “Slide Synchronization Properties” feature of DrawingML provides the ability for a presentation to synchronize slide content in centrally stored slides on a server. This is known to be a feature of Microsoft PowerPoint and SharePoint. However, nothing in the description of this feature allows anyone but Microsoft to interoperate.

Although sufficient standards exist for describing a client-server protocol of this sort, namely the various Web Services standards, OOXML gives us no information.

This record in the file format is declared but not defined and will be of no use to anyone but Microsoft.

Interoperable & Portable (Summary)

Our observation is that wherever OOXML has reference to other technologies it has been designed to work exclusively with the technologies supported by Microsoft Office. In some cases extraordinary efforts are made to incorporate other specifications, like VML, into OOXML. Not only does OOXML ignore alternative, standard and open technologies, it prevents other vendors from adding this support.

Of course, Microsoft is entitled to their own design decisions and their own priorities. No one denies this. However, an ISO standard must have the characteristics of portability and interoperability, so other vendors may have that same right to their own design decisions and priorities.

Examples

- Data interchange between spreadsheets and RDBMSs
- Clipboard Formats
- Password Hashing
- Optimize for Browser
- Slide Synchronization Properties

The arbitrary restrictions of OOXML, which fit Microsoft like a glove, but suit the rest of us so poorly, render the proposed standard inappropriate for approval as an International Standard.

Cultural & Linguistic Adaptability

Since OOXML's feature intentionally maps directly onto the feature set of Microsoft Office, it is not surprising that this feature set best reflects the needs of developed countries and communities where Microsoft's business has seen the greatest success. However, an International Standard must take a broader view and provide broad cultural and linguistic interoperability.

Cultural & Linguistic Adaptability (I)

“NETWORKDAYS()”

An example of a concern is the spreadsheet function NETWORKDAYS(). This function is defined by OOXML to return the number of working days between two dates, exclusive of any weekends in that interval. In Western Europe and North America, the weekend is Saturday and Sunday. In the Middle East, however, the days of rest are either Thursday/Friday or Friday/Saturday.

OOXML does not define “weekend” and does not provide a way for the user to define it either. As implemented in Excel the function assumes the weekend is always Saturday/Sunday.

So this spreadsheet function is defined in a way which renders an incorrect answer for over one billion people. It lacks cultural adaptability.

Compare this to the same function in OpenDocument Format, where the user may pass in an additional parameter to override the default definition of a weekend.

“Numeration Styles”



As mentioned previously, WordProcessingML defines a number of numeration styles for numbered lists. There we pointed out that these numeration styles were essentially only labeled but not defined. Here we also raise the issue that these styles are defined as a closed list, again matching what Microsoft Word supports, but is not extensible by other vendors. Since the list of styles provided is incomplete, lack, for example, support for Armenian, Tamil, Greek alphabetic, Ethiopic and Khmer numerations, as well as the larger number of historic systems used by scholars.

The preferred solution is to use a declarative/generative approach, such as used by the W3C's XSL:FO and OpenDocument Format. This allows an open-ended list of numeration styles to be used, each self-defining.

Cultural & Linguistic Adaptability (I)

“Border Styles”

WordProcessingML has a feature called “Border Styles”² which lists a large number of graphical borders which can be used as page borders. These represent a closed list of specific named border styles with mandated images. An example of two such graphics is shown in figure 1. Note that these are the only two possibilities for displaying a globe in a page border and neither of them show Asia.

earth1 (Earth Art Border)	Specifies an art border consisting of a repeated image of Earth, as follows (showing two repetitions): 
earth2 (Earth Art Border)	Specifies an art border consisting of a repeated image of Earth, as follows (showing two repetitions): 

Similarly, there are graphics for birthday cakes, St. Valentine's Day cupids, painted Easter eggs, Christmas gingerbread men, Halloween Jack O'Lanterns, and other images that are appropriate for a Western cultural milieu, but have limited application elsewhere. The primary problem here is that this list of page border styles is a closed list. A vendor cannot extend this list with additional images types to better suit the cultural milieu of their customers. It matches exactly what Microsoft Word provides, but does not give a facility for a vendor to improve on or extend this list if they determine, for example, that having a picture of the earth with Asia appearing on it is desired for their product and their market. This is what is meant by “cultural adaptability”. How well does OOXML adapt to other cultures? In the case of page borders we note that it fails to adapt.

Cultural & Linguistic Adaptability (Summary)

The theme is that cultural and linguistic adaptability suffers in OOXML because of closed-ended lists which, although they may match perfectly what Microsoft Office offers today, are not extensible by vendors in an interoperable way.

Examples

- NETWORKDAYS()
- Numeration Styles
- Border Styles

Summary

Standards have standards. In this presentation we have evaluated the proposed OOXML standard based on the criteria provided by ISO for what a standard should be. We looked at it from the perspective of the various desired characteristics of ISO standards:

precision, common criteria, optimal degree of order, aspirational, consolidation of the best practices of science, technology and experience, interoperability, portability and cultural and linguistic adaptability.

We have provided many examples of where the proposed OOXML standard has fallen short of this mark. By failing to meet these criteria OOXML has failed to provide for the optimum community benefit. Indeed, the proposal appears to be targeted to benefit a single corporation only.

Summary

The expectations for a document format standard are high, and they should be. A standard document format that meets the above criteria is essential to long-term preservation of our digital heritage, for equal access to government documents and records by all citizens, and for cost-effective and efficient document-based business process integration and workflows across heterogeneous systems.

Microsoft OOXML, the file format for Microsoft Office, does not provide these benefits, and is not suitable for an ISO standard, and therefore we urge a vote of disapproval in this JTC1 ballot.