

# Office OpenXML

May 2007

Adam Farquhar

## Outline

- Office OpenXML
- Importance to Library and Archive community
- History
- Relation to other standards
- Design criteria
- Structure of standard
- Working with the specification
- Conclusion

## Office OpenXML – A standard file format for Office Documents

Office OpenXML is an open standard for word-processing documents, presentations, and spreadsheets

High Fidelity Migration from legacy Microsoft binary formats

- Faithfully represent in XML the pre-existing corpus of word-processing, presentations and spreadsheets documents
- Millions of users created billions of documents over the past 20 years

Interoperability, Platform independence, Internationalization, Accessibility

- Extensive review and modifications during the standardization process

Enable new range of applications - Integration with business data

- Clear definition of conformance
- Support Custom XML Schemas (e.g. Birth Certificate, HL7)

Long-term preservation

- Full specification, no application or system dependencies, clear path for migration, future evolution/maintenance in Ecma & ISO

## What is wrong with legacy binary formats?

- Designed to be manipulated by a single vendor's software
- Direct serialisation of in-memory data structures
- Evolved over many years in response to customer needs
- Augmented through acquisitions
- New features re-used existing attributes
- Result – the software is the specification!  
“We are renting our content from Microsoft”

## Why should libraries or archives be involved?

Address a root cause of digital obsolescence

- Formats have been deeply coupled with the programs that create them
- Formats are often poorly specified and complex
- Programs have a shorter lifespan than content

Raise awareness about digital preservation

- Especially among software vendors
- The standard identifies preservation as a key issue

Represent our interests

Provide an independent voice

Save pain down the line

- Compare bulk de-acidification, treating caustic inks

## Office Open XML - History

### Start

- In 2000, Microsoft became serious about using XML in its Office file formats
- Consumers, governments, libraries, archives become increasingly vocal about the need for a full specification for the Office file formats
- Microsoft Office 2003
  - XML formats published on Danish Government site
  - IDA (2004) <http://europa.eu.int/idabc/en/document/2592/5588>  
“Microsoft should consider the merits of submitting XML formats to an international standards body of their choice”
  - IDA & EC explicitly ask Microsoft
    - to put the evolution of the formats in the control of a standards body
    - to build translators to/from ODF
  - Governments recommend eventual submission to ISO

### Now

- Dec 2006 - PEGSCO Report
  - Microsoft has adopted a “pure” XML format
  - The Open XML (ECMA-376) standard is freely available
- The Open Specification Promise enables both Open Source and Commercial software to implement Office OpenXML

## Ecma-376 Office Open XML Standardization

November 15 2005, co-submission of Office Open XML Formats to Ecma International

Co-sponsors: Apple, Barclays Capital, BP, British Library, Essilor, Intel Corporation, Microsoft Corporation, NextPage Inc., Statoil ASA, Toshiba

- Participants represented a wide range of interests

December 8 2005, Ecma General Assembly accepts standardization: Ecma TC 45 created

Goal:

- To create an Ecma Office OpenXML Format standard
- To contribute the Ecma Office Open XML Format standard to ISO/IEC JTC 1 for approval and adoption by ISO and IEC
- To steward future evolution of Office OpenXML

Open process

- Technical Committee open to any Ecma member
- Novell, US Library of Congress joined TC45 after creation

## Ecma Standardization

- Dec15, 2005 - 1st face to face meeting – Brussels
- Microsoft submit initial 2000 page draft of Office Open XML
- Weekly 2 hour conference call – 15-20 participants
- Face 2 face @ Ecma, Apple, British Lib, Toshiba, Microsoft, Statoil
- Initial and Interim drafts posted publicly on Ecma web site
- External feedback – SC34 experts, others
- Final standard 6000 pages
- Ecma GA: Overwhelming positive vote - Approval to Submit to ISO



**Ecma Secretary General  
Jan van den Beld (left)  
receives initial draft of  
office document standard  
from TC45 Chair Jean Paoli  
(center)**

**Adam Farquhar (right),  
TC45 Vice-Chair,  
Head of e-Architecture for  
the British Library**

## ECMA Process

- ECMA has an established well-defined open standards process
  - Any ECMA member was open to joining the committee
  - ECMA membership is open
  - Government and academic membership is free
  - My observation: very professional organisation
- TC45 took openness seriously
  - Published progress reports after each face-to-face meeting
  - Published multiple interim versions of the evolving specification
  - Accepted email comments and feedback
    - These were added to the issue log and treated with the same consideration as issues raised by committee members
- TC45 also took progress seriously!
  - Substantial initial specification
  - Clear timeline
  - Dedicated editor, support staff

## Ecma-376 Office Open XML Adoption Many Office suites - Multiple platforms

Microsoft Office 2007 - Default Save Format is Open XML (+ free updates for Office 2000, XP, 2003) – Dec 2006/Jan 2007

Open Office – Novell support Open XML in Open Office – Novell edition – Availability Feb 2007

Corel – announcement of support of Open XML - Availability mid 2007

Gnumeric – open source Spreadsheet supports OpenXML

Sun – working on OpenXML import filter for spreadsheets

OpenXMLDeveloper.org (hundred of developers, multiple platforms)

## ISO Standardization

### Ecma General Assembly approval

- Dec 2006 - Overwhelming Positive vote for approving sending Open XML to JTC1 ISO Fast Track

### ISO Fast-Track Process

- JTC1 Fast Track procedure - Approved for Ecma Standards
- >75% of Ecma standards approved as ISO/IEC standards

### Ballot time

- Jan 5 – Ecma submit Office Open XML to ISO/JTC1
- Feb 5 – End of 30-day review period, to determine perceived contradictions
- Feb 28 – Ecma provides feedback on comments & perceived contradictions
- 5-month letter ballot – Technical Review through September 2<sup>nd</sup>

## The Highlander myth

How many document format standards should there be?

Some say they can be only one (The Highlander Principle)

- As sensible as the movie! Where otherwise immortals slay each other!

In fact, there are many standard formats now:

- HTML, PDF/A, ODF, OOXML
- CGM, SVG; JPEG, PNG; TIFF/IT, PDF/X
- And many more widely used formats

And there will continue to be many

- No format is immortal
- Formats address different needs
- Innovation is not over

## The simple office document myth

Have you heard – office documents are simple!

- In fact, they can be extraordinarily complex
- Office documents can contain:
  - Multiple character sets
  - Left-to-right, right-to-left, bi-directional text
  - Images, sound, video, vector graphics
  - Annotations and changes from multiple authors
  - Arbitrary metadata and XML components
  - Complex mathematical equations
  - Animated transitions
  - Embedded data, database connections, queries, cached data
  - Embedded components from other applications

## The monolithic specification myth and the proportionality principle

Six thousand pages! That's too big for anyone to use.

In fact, the standard follows a proportionality principle

Easy jobs should be easy!

- A developer can take the standard and implement tools within a week (assuming knowledge of zip, xml)
- Examples: update email addresses or copyright notices, replace logos, extract text stream, produce simple documents

Hard jobs can be hard!

- A implementing a full office suite will take many person-years
- Examples: provide high-performance calculation engine, provide full OOXML->ODF translation, develop an MS Office competitor
- But all of these are now possible!

## The ECMA-376 Specification

The committee worked to make it readable!

White Paper (14p)

Part 1: Fundamentals (165p)

- Accessible with simple examples

Part 2: Open Packaging Conventions (125p)

Part 3: Primer (466p)

- Many examples, diagrams, explanations

Part 4: Mark-up Language specification (5756p)

- Detailed, but most uses require only small subsets

Part 5: Compatibility and extensions (34p)

Extensive cross-references ease navigation, but add bulk

# Open XML Format Architecture



User view: single document

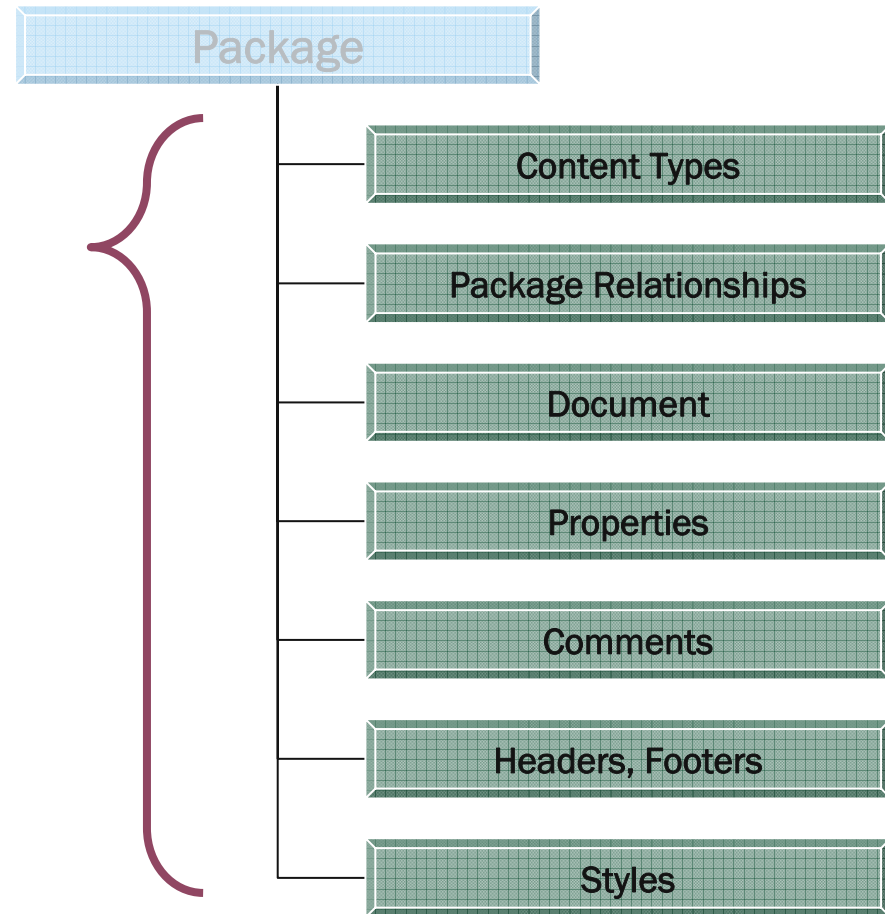


Sample.docx

Developer view: modular file

## Document Parts

- Most parts are XML
- Each XML part is a discrete, optionally compressed component
- Can add, extract and modify individual parts without using Office programs
- Corruption or absence of any part does not prohibit the file from being opened



## OpenXML Mark-up approach

- Very different mark-up approach from ODF, HTML
  - Flatter structure
  - Local edits result in local changes

Basis for text is a **run**

- A run is contiguous text with identical properties  
This is **three** runs

```
<w:p>  
  <w:r><w:t xml:space="preserve">This is </w:t></w:r>  
  <w:r>  
    <w:rPr>  
      <w:b /><w:color w:val="00CCFF" />  
    </w:rPr>  
    <w:t>three</w:t>  
  </w:r>  
  <w:r><w:t xml:space="preserve"> runs .</w:t></w:r>  
</w:p>
```

# The principle of proportionality confirmed

## Open XML Demo

You need to fill a wordprocessingML page for demo purpose, use the lorem function → =lorem(5) ©

Consectetuer adipiscing elit. Maecenas porttitor congue massa. Fusce posuere, magna sed pulvinar ultricies, purus lectus malesuada libero, sit amet commodo magna eros quis urna.



Nunc viverra imperdiet enim. Fusce est. Vivamus a tellus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames

ac turpis egestas. Proin phare nonummy pede. Mauris et or

Aenean nec lorem. In porttitor laoreet nonummy augue.

Suspendisse dui purus, sceler vulputate vitae, pretium mat Mauris eget neque at sem vel eleifend. Ut nonumm

Contact	Company
 <p>Julien Chable Wygwam France 1 rue de la Performance – F-59650 Villeneuve d'Ascq <a href="http://www.wygwam.com/">http://www.wygwam.com/</a></p>	
Mail : julien@wygwam.com	
Tél : +33 (0)3 20 82 38 77 Mobile : +33 (0)6 74 28 40 50	
myemail@world.com azerty@azerty.com	

```
Problems Declaration Search Console Progress Hi
<terminated> DemoExtractEmails (1) [Java Application] C:\Progra
3 mail(s) found ! Here is the list :
julien@wygwam.com
myemail@world.com
azerty@azerty.com
```

- New open source project from Julien Chable
- Bulk of code serves to manipulate packages
- A few minutes sufficed to write a tool to extract email addresses from any OOXML document

## Conclusion

The Digital Library community has influenced Office OpenXML

- Key vendors are more aware of digital preservation

The Office OpenXML Standard

- Co-exists with existing and future document standards
- Plays a key role preserving billions of legacy office documents
- Follows the Proportionality Principle
- Enables innovation
- Is in the ISO process
- Continues to evolve through an open process

Now we own our content!

Questions?