XML For The Massses
An Open Office XML File Format

by

Michael Brauer
mib@openoffice.org

Agenda

- About the speaker
- Introduction into OpenOffice.org XML Format
- Detailed View
- OASIS/OASIS Open Office XML Format TC
- Conclusion/Questions

About The Speaker

- Michael Brauer
- Technical Lead at Sun Microsystems, Inc.
- OpenOffice.org XML Project Owner
- OASIS Open Office XML Format Technical Committee chair

What's Wrong With Your Office

The Case for an Open Office XML Format

- Voice of the Customer
  - office productivity applications
  - need additional processing, integration
    - Archiving and indexing
    - content checking
    - database, work-flow integration
    - Operation on many files
  - need long-term readability
  - want application independence
  - unstructured work-flow, many document types
- Requirement
  - Preserve functionality, but open file format

XML Office File Format

Matching Customer's Requirements

- Solution: Office XML File Format
  - XML, for easy integration, processing
  - define office vocabulary
  - requirements for office file format
    - full featured, cover full office productivity space
    - easy to process, easy to generate
    - not vendor or application specific
- OpenOffice.org XML File Format
  - tailored to these requirements

Fixed vs Custom Vocabularies

Why the World Needs an Open Office XML Format

- Custom Vocabularies
  - requires tooling, extensive preparations
  - work well in specialized, structured work-flows
- Fixed Office Vocabulary
  - allows traditional usage patterns
  - mass-market compatible
  - add XML processing as needed
  - tools can operate on semantic units
  - transform into custom vocabularies

Standardizing The Format

Securing The User's Investment

- A File Format is an Investment
  - must be open, documented
  - must be stable, controlled evolution
- Win-Win with Widespread Adoption
  - more support, more tools
  - need user and industry acceptance
- Develop Standard Format at OASIS
  - OASIS Open Office XML Format
  - based on OpenOffice.org XML Format
    - format has proven useful in real life

- Format Requirements
  - 1st class XML
  - easy transforms
- Format Details
  - content vs layout
  - binary data

Designing a File Format

- Meeting the Requirements
  - existing formats? Not sufficient.
  - embrace and extend? No!
  - XML-ify existing structures? No!
- Reviewed Design Process
  - examine existing formats
    - use MathML, XLink, Dublin Core
    - reuse from XHTML, SVG, XSL-FO, CSS
  - specify, review, finally implement

First Class XML

Map all Structured Content to XML

- Fully Compliant: XML, namespaces
- Use XML for Structured Content
  - no information in physical representation
  - no information in comments
  - no information in 'special' names or values
  - no 'sub-formats' for values
- Values Make Sense
  - values vs. presentation
  - process what you need

2002-12-10T11-45
December 12th, 2002
37541.49

Easy Transformations

Making it Simple to Access Office Files

- Consistent Design
  - common format across all applications
  - one concept, one representation
- Reuse of Vocabularies
  - HTML, SVG, DC, MathML, XLink, XSL
- Examples
  - all text in <text:p>, <text:h>
  - extract plain text: 2 XSLT rules
    - add more as you go along (+4 for footnotes)

Styles – Content vs Layout

"Markup reflects a theory of text."
C. M. Sperberg-McQueen

- Office World
  - document = content + layout
  - Layout part of user input
- Semantic Markup
  - document = content
  - layout separate, external (CSS, XSLT)
- We Keep Both, Separately

Styles – Content vs Layout

- Styles Separate From Content
  - easy to change layout
  - easy to process
  - transparent to user
- Style Section(s)
  - convert all formatting into styles
  - separate container elements
  - built-in stylesheet

```
<style name="Emphasis">
    <properties
    font-weight="bold"/>
</style>

<p style-name="Emphasis">
    text text text
</p>
```

Packages

Efficient Handling of XML and Binary Data

- XML File Format Concerns
  - file size
  - embedded images, objects
- XML Package
  - ZIP format, XML-based manifest
  - XML streams + binary streams
    - images, OLE objects, printer setup data
  - used by OpenOffice.org, Gnome, KOffice
- 'pure' XML: embedded BASE64

About OASIS

- OASIS = Organization for the Advancement of Structured Information Standards
  - Non-Profit, global consortium for e-business standards
    - Web services
    - business transactions
    - electronic publishing
    - interoperability within and between marketplaces
    - etc.
  - Lightweight and open processes
    - expressly designed to promote industry consensus and unite disparate efforts
  - More than 600 corporate and individual members
    - Sun Microsystems
    - Corel
    - Oracle
    - Microsoft
    - etc.
  - Low membership fee
    - Starts with 250 USD for individual members

OASIS Open Office XML TC

- Members
  - Office Application Vendors
    - Sun Microsystems
    - Corel
    - KOffice
  - XML Tool Vendors
    - Arbortext
    - Corel
    - Stellent
  - Industry and Users Groups
    - Boeing
    - National Archive of Australia
    - Society of Biblical Literature
    - SpeedLegal

OASIS Open Office XML TC

The Committee's Work

- Committee started Dec., the 16th, 2002
  - First phase: Base specification for text documents, spreadsheets, drawings, presentations and business charts
  - Second phase: Extended applications
- Weekly conference calls
- Face-To-Face Meeting Feb., the 18th/19, 2003
  - Menlo Park, California
- Committee Specification planned for Summer 2003 (End of phase 1)
- Planned to become OASIS Specification
- Topics decided so far
  - Overall structure
  - Meta information
  - Extensibility

Conclusion

- Office XML File Format
  - fully supports office users & applications
  - Enables
    - processing of office documents
    - integration into custom infrastructure
  - fixed vocabulary is key
- Standardization at OASIS
  - open standard, open development
  - secures your investment

Links & questions

- OpenOffice.org http://www.openoffice.org
  - XML Project http://xml.openoffice.org
- OASIS http://www.oasis-open.org
  - Open Office TC http://www.oasis-open.org
  - Open Office TC Mailing List http://www.oasis-open.org/committees/office/
  - Public Comment Open Office TC Mailing List http://office-comment@lists.oasis-open.org

Questions?