THE OASIS OPENDOCUMENT FORMAT

Michael Brauer
Technical Architect Software Engineering
Sun Microsystems GmbH

About the Speaker

- Technical Architect in OpenOffice.org/StarOffice development
- OpenOffice.org/StarOffice developer since 1995
- Main focus: XML technologies
- OpenOffice.org XML project owner
- OASIS OpenDocument Technical Committee chair

Agenda

- History of OASIS OpenDocument format
- Overview
- Technical Details

History

History

- 1999
  - Development of StarOffice XML file format starts
    - Primary goal: interoperability
- 2000
  - Foundation of OpenOffice.org and OpenOffice.org XML project
    - Sun contributes StarOffice to OpenOffice.org
    - Name change to “OpenOffice.org XML”
    - OpenOffice.org XML becomes open source community project
  - First OpenOffice.org XML working draft available

History

- 2001
  - 1st OpenOffice.org XML implementation (OpenOffice.org 1.0 and StarOffice 6.0)
    - Used as default file format, no alternatives
    - Used by all applications
    - Proof of concept
- 2002
  - OpenOffice.org/KOffice collaboration on package format
  - Foundation of OASIS Open Office Technical Committee
    - Goal: vendor independence
    - Basis of TC's work: OpenOffice.org format

History

- OASIS Open Office XML charter (extract)
  - The purpose of this TC is to create an open, XML-based file format specification for office applications.
  - The resulting file format must meet the following requirements:
    - it must be suitable for office documents containing text, spreadsheets, charts, and graphical documents,
    - it must retain high-level information suitable for editing the document,
    - it must be friendly to transformations using XSLT or similar XML-based languages or tools,
    - it should 'borrow' from similar, existing standards wherever possible and permitted.

History

- 2004
  - OASIS Open Office XML Committee Draft 1 and 2
- 2005
  - Name change to “OASIS Open Document Format for Office Applications (OpenDocument)”
    - Emphasizes application and use case independence
  - OASIS OpenDocument v1.0 gets approved OASIS standard

- 2005
  - (OASIS) OpenDocument submitted to ISO

History

- Changes from OpenOffice.org to OpenDocument
  - New XML Namespaces
  - Schema language: Relax-NG (OOo: DTD)
  - Schema/Validation improvements
    - Elements for document classes, etc.
    - Attribute value data types (XSD)
  - Adoptions to new standard versions
    - SVG 1.1, CSS3, etc.
  - Adoption to new/other office applications
  - Removal of inconsistencies, error corrections

Overview

Basics

- Office versus Custom Schema
  - Office Schema: Schema for office documents
    - Tailored to office documents
    - Feature complete
  - Custom Schema: Any other schema
    - Standard schemas: UBL, XHTML, DocBook, etc.
    - User defined schemas: Defined by customer
      - Bills, orders, meeting minutes, etc.

Basics

- Office Schema
  - "Own your data" for office documents
    - Interoperability
    - Long term preservation
  - Loading and saving of office documents
  - Default file format

Completeness

- Completeness
  - Regarding document types (text, spreadsheet, etc.)
  - Regarding features of document types
- Why is this important?
  - Long term preservation for *all* office documents
  - Key requirement for interoperability
  - Precondition for being default file format
- OpenDocument schema
  - Covers text, spreadsheet, graphics, presentations, business charts and formulas

Interoperability

- OpenDocument Schema
  - Interoperability explicit goal
  - Designed independently of OpenOffice.org application
  - Proof of concept by KOffice adoption
  - Reuses standards
    - XHTML, SVG, SMIL, XSL, XForms, MathML, XLink and Dublin Core Meta Initiative
    - Reuses its own schema fragments

Default File Format

- Default file format - Why is this important?
  - Ensures *all* data created can be processed
  - Avoids use of historic binary formats
  - Documents can be reused
    - By other office applications
    - For tasks not yet known
- OpenDocument Schema
  - Usable as default file format
    - Default file format of OpenOffice.org 2.0/StarOffice 8
    - OpenOffice.org XML format is default file format of OpenOffice.org 1.0/StarOffice 6.0 since 2001

File Size

- File Size - Why is this important?
  - Office Documents may get huge
    - 800 page design specification, 1,000,000 cell business plan
  - File size user experience with binary formats
  - Transfer of documents
  - Bandwidth for e-mail transfer
  - Storage requirements for archiving
- OpenDocument Schema
  - Package concept
    - Small file size through compression
    - Document and images in one physical file

Technical Details

Package Concept

- Structure
  - OpenDocument packages are ZIP files
  - XML manifest file for mime types/encryption details
  - References
    - To targets within same package: relative URIs
    - To targets outside packages: absolute or relative URIs
    - To other packages: not available (so far)
    - Relative URIs are resolved by “virtually” extracting package to a folder with the package name

Package Concept

- Content
  - “meta.xml”
    - Office document meta data
  - “styles.xml”
    - Styles (page styles, paragraph styles, etc.)
    - “style” = style as available in office applications
  - “content.xml”
    - Content (paragraphs, lists, tables, etc.)
    - Automatic styles
    - “automatic style” = style generated for hard formatting
  - “settings.xml”
    - Application specific settings (visible area, cursor position, etc.)

Package Concept

- Content
  - Embedded objects
    - OpenDocument objects: sub folders, same structure as top level documents
    - Other objects: binary streams
  - Images
    - PNG, JPG, etc.
  - Non standardized content
    - Macros
    - Menu and toolbar configurations
    - Arbitrary other streams and folders

Design Principles

- One Concept = One Schema Fragment
  - Paragraphs: Same schema in
    - Text documents,
    - Spreadsheet table cells
    - Presentation graphical shapes
  - Tables: Same schema in
    - Text documents
    - Spreadsheets Tables
  - Graphical Shapes, Lists: Same schema in
    - ...

Example

OpenDocument Text Table

```
<table:table
      table:name="Table1"
      table:style-name="Table1">
  <table:table-column
      table:style-name="Table1.A"/>
  <table:table-row
      table:style-name="Table1.1">
    <table:table-cell
        table:style-name=
            "Table1.A2"
        table:value-type=”string”>
      <text:p>Hello
              World!</text:p>
    </table:table-cell>
  </table:table-row>
</table:table>
```

OpenDocument Spreadsheet Table

```
<table:table
      table:name="Table1"
      table:style-name="Table1">
  <table:table-column
      table:style-name="Table1.A"/>
  <table:table-row
      table:style-name="Table1.1">
    <table:table-cell
        table:style-name=
            "Table1.A2"
        table:value-type=”string”>
      <text:p>Hello
              Word!</text:p>
    </table:table-cell>
  </table:table-row>
</table:table>
```

Design Principles

- Document Type = Composition of Fragments
  - Text Document
    - Sequence of paragraphs, lists and tables
    - Paragraphs may contain graphical shapes
  - Spreadsheet
    - Sequence of tables
    - Table cells contain sequence of paragraphs
  - Presentation
    - Sequence of Pages
    - Pages contain sequence of graphical shapes

Example

OpenDocument Text

```
<office:document-content>
  <office:automatic-styles/>
  <office:body>
    <office:text>
      <text:p/>
      <text:p/>
      <table:table>
        <table:table-row>
          <table:table-cell>
            <text:p/>
          </table:table-cell>
        </table:table-row>
      </table:table>
    </office:text>
  </office:body>
</office:document-content>
```

OpenDocument Spreadsheet

```
<office:document-content>
  <office:automatic-styles/>
  <office:body>
    <office:spreadsheet>
      <table:table>
        <table:table-row>
          <table:table-cell>
            <text:p/>
          </table:table-cell>
        </table:table-row>
      </table:table>
      <table:table/>
      <table:table/>
    </office:text>
  </office:body>
</office:document-content>
```

Design Principles

- Values and Presentation
  - Values are of interest for office applications
  - Presentation is of interest for viewers, transformations, etc.
  - OpenDocument usually stores both
  - Example (value, presentation):

```
<table:table-cell office:value-type="float"
                  office:value="42">
  <text:p>42,00</text:p></table:table-cell>
<table:table-cell table:formula="oooc:=[.A1]/7"
       office:value-type="float" office:value="6">
  <text:p>6,00</text:p></table:table-cell>
```

W3C Recommendations

- Included:
  - XML 1.0
  - XML Namespaces 1.0
  - XLink 1.0
  - XForms 1.0 (partially)
  - MathML 2.0

W3C Recommendations

- Based on:
  - “based on” = OpenDocument uses attributes with same name and semantic as a W3C Recommendation, but within its own namespace
  - XSL-FO 1.0
  - SVG 1.1 (elements and attributes)
  - SMIL 2.0
- Other:
  - HTML4 (paragraph, list and table structure equals HTML4)
  - CSS2/CSS3 (property names and values)

Example (Inclusion)

```
<office:document-content
 xmlns:office="urn:oasis:names:tc:opendocument:xml
 ns:office:1.0"
 xmlns:xlink="http://www.w3.org/1999/xlink" [...]>
 <office:body>
  <office:text>
   <text:p text:style-name="Standard">
      <text:a xlink:href="http://myUrl">Hello
        World!</text:a>
   </text:p>
  </office:text>
 </office:body>
</office:document-content>
```

Example (Based on)

```
<office:document-content
 xmlns:fo="urn:oasis:names:tc:opendocument:xmlns:xsl
 -fo-compatible:1.0" [...]>
 <office:automatic-styles>
  <style:style style:name="T1" style:family="text">
   <style:text-properties fo:font-weight="bold"/>
  </style:style></office:automatic-styles>
 <office:body>
  <office:text>
   <text:p text:style-name="Standard">Hello
 <text:span text:style-
 name="T1">World</text:span>!</text:p>
  </office:text>
 </office:body></office:document-content>
```

Example (Based on)

```
<office:document-content
 xmlns:svg="urn:oasis:names:tc:opendocument:xmlns:
 svg-compatible:1.0" [...]>
    [...]
    <draw:page>
      <draw:rect draw:style-name="gr1"
           svg:width="6cm" svg:height="3cm"
           svg:x="3.822cm" svg:y="3.008cm">
        <text:p/></draw:rect>
      <draw:ellipse draw:style-name="gr1"
           svg:width="5cm" svg:height="2.5cm"
           svg:x="7.822cm" svg:y="7.008cm">
        <text:p/></draw:ellipse>
      </draw:page> [...]
```

Questions & Answers

Resources

- OASIS
  - http://www.oasis-open.org
- OASIS OpenDocument Technical Committee
  - http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=office

THE OASIS OPENDOCUMENT FORMAT

Michael Brauer
michael.brauer@sun.com