Document Formats for Archiving

James King
Leonard Rosenthal
Diana Helander
Adobe Systems Incorporated
Outline

- Development Stages and Document Formats
- Format Properties
- Document Archiving Requirements
- XML Myths
- Conclusions
Development Stages and Document Formats
Stages of Document Development

Template Processor → Layout Formatter → Device Renderer

Display → Print
Document Formats Before Binding Stages

Stage 1

Least Bound

Template Processor

Stage 2

Layout Formatter

Stage 3

Device Renderer

Stage 4

Display

Print

Most Bound
Document/Data Type at Each Stage

Text Streams
- e.g., pictures, fonts

Style Directives

Template Processor

Flowable Presentation

Stage 1

Resources

Layout Formatter

Final Presentation

Stage 2

Device Renderer

Stage 3

Display

Print

Stage 4
Example Document Formats at Stage 2 and Stage 3

- **Stage 1**: Text Streams, Style Directives, Resources (e.g., pictures, fonts)
  - Template Processor

- **Stage 2**: Flowable Presentation
  - Layout Formatter
  - Final Presentation

- **Stage 3**: Device Renderer
  - Display
  - Print (e.g., Scanned Pages PDF, PDF/A)
  - E.g., OOXML ODF

- **Stage 4**: Final Presentation
  - E.g., Converted Formats (e.g., PDF, PDF/A)
Flowable/Editable Document Formats
Native Document Format for the Application

Text Streams
- e.g., pictures, fonts

Style Directives

Template Processor

Resources

Flowable Presentation

ODF

Writer For ODF

Open Office.org

Stage 1

Stage 2

Stage 3

Stage 4

Device Renderer

Display

Print

e.g., OOXML
ODF

e.g., Scanned Pages
PDF, PDF/A
Writer Produces PDF or PDF/A
Stage 2 to Stage 3

Text Streams
Style Directives
Resources e.g., pictures, fonts
Template Processor
Flowable Presentation
Open Office.org
Writer For ODF
PDF PDF/A
Device Renderer
Display
Print
Stage 1
Stage 2
Stage 3
Stage 4

e.g., OOXML ODF

e.g., Scanned Pages PDF, PDF/A

e.g., OOXML ODF

For ODF

Flowable Presentation

Final Presentation

Writer Produces PDF or PDF/A
Stage 2 to Stage 3

Text Streams
Style Directives
Resources e.g., pictures, fonts
Template Processor
Flowable Presentation
Open Office.org
Writer For ODF
PDF PDF/A
Device Renderer
Display
Print
Stage 1
Stage 2
Stage 3
Stage 4

e.g., OOXML ODF

e.g., Scanned Pages PDF, PDF/A

e.g., OOXML ODF
Rendered by Adobe Reader and Other Products
Stage 3 to Stage 4

Text Streams
- e.g., pictures, fonts

Style Directives
- Flowable Presentation

Template Processor

Resources
- e.g., OOXML
- ODF

Stage 1

Stage 2

Writer For ODF

Stage 3

Adobe Reader

Stage 4

Display

Print

e.g., Scanned Pages
- PDF, PDF/A

e.g., OOXML
- ODF

Open Office.org

PDF
- PDF/A

Resources
- e.g., Scanned Pages
- PDF, PDF/A
Displayed or Printed
Stage 4

Text Streams
e.g., pictures, fonts
Style Directives

Resources

Template Processor

Flowable Presentation
e.g., OOXML, ODF

Open Office.org

Writer For ODF

PDF, PDF/A

Adobe Reader

Display
Print

Stage 1
Stage 2
Stage 3
Stage 4
Format Properties
Reusing Document Content

Text Streams
- Resources: e.g., pictures, fonts
- Style Directives

Stage 1
- Template Processor
- Flowable Presentation

Stage 2
- Layout Formatter
- Final Presentation

Stage 3
- Device Renderer
- Ease of reuse

Stage 4
- Display
- Print

Resources
- e.g., OOXML, ODF
- e.g., Scanned Pages, PDF, PDF/A

Ease of reuse
Preserving Authors Design

Stage 1
- Text Streams
- Style Directives
- Resources (e.g., pictures, fonts)

Stage 2
- Template Processor
- Flowable Presentation
- Layout Formatter

Stage 3
- Final Presentation
- Appearance Fidelity

Stage 4
- Display
- Print
- e.g., Scanned Pages (PDF, PDF/A)
- e.g., OOXML (ODF)
Document Archiving Requirements
Document Archiving Requirements: ePaper

- Paper, unchanged but in electronic form
- Preserve public information
  - e.g., birth certificates, wedding licenses
- Business transactions
  - e.g., invoices, agreements
- Laws and Legal decisions
  - e.g., records from legislative bodies, court records
- Objective
  - Preserve for future reference as if paper documents – emphasizing visual fidelity
- May have been scanned paper documents
- Edit and reuse is not a requirement
ePaper

Stage 1

Text Streams
- e.g., pictures, fonts

Style Directives

Template Processor

Resources
- e.g., OOOXML, ODF

Stage 2

Flowable Presentation

Layout Formatter

Stage 3

Final Presentation

Device Renderer

Stage 4

e.g., Scanned Pages
- PDF, PDF/A

Display

Print

Copyrigh 2008 Adobe Systems Incorporated. All rights reserved.
Document Archiving Requirements: Editable

- An editable master document
  - edited to produce derivative documents
  - over long periods of time

- Format primarily designed to preserve the edit-ability of the contents
- Sacrificing exact appearance OK in order to provide edit-ability
- Derivative document appearance manipulated when derived
Document Archiving Requirements: Pictures aka Images

- Images consume storage
  - Compression important

- Metadata

- Color images
  - Compress much more if lossy compression is allowed
  - The best lossy color compression remove image features that are impossible for humans to physically see under normal viewing conditions.

- Archiving requirements confined to normal human viewing of the archived pictures then
  - Careful application of lossy color compression is acceptable.

- (show a PDF file compose primarily of pictures)
Document Archiving Requirements: Scanned

- **ePaper**
  - Potentially superior to paper for archiving
  - Gives low cost access to everyone

- **TIFF is typical choice**
  - TIFF technology includes RGB and CMYK color representations and JPEG, LZW, fax, …

- **Often used with no compression**
  - Easily retrieved, no accidental loss (e.g., error in compression software)

- **Risks**
  - Not a publicly managed standard (owned by Adobe)
  - Incompatibly fragmented (TIFF 6 published in 1992)

- **Need open standards – then use PDF/A**
  - Compression choices: none, fax, flate (PNG), JBIG2 and JPEG.
  - Optical character recognition and document recognition resulting in non-image PDF/A
    - Enables word searching & indexing
  - Well defined and extensible metadata (XMP)
Image Filters within PDF 1.7

- **B/W pictures**
  - Flate
    - general purpose byte compression (page contents)

- **Color pictures**
  - DCT (JPEG)
    - usually lossy but good for color pictures
  - JPX (JPEG2000)
    - uses wavelets instead of discrete cosine transforms (not in PDF/A)

- **Scanned text**
  - CCITTFax
    - G3 and G4 fax standards
  - JBIG2
    - pattern compression good for text

- **Others**
  - ASCIIHex
    - to have printable characters
  - RunLength
    - compatible with Unix compress
  - Crypt
    - encryption technology (not in PDF/A)
Document Archiving Requirements: Live

- interactive and dynamic document types
- more powerful computers and displays available
- invent new document metaphors that move away from our more static paper-based ideas
- two-fold challenge
  - obsolescence of new document types
  - base technologies may become obsolete
- Live documents
  - games and interactive web applications.
- repeat the current experience at some future time.
  - active nature and that depends upon being able to provide a suitable “execution environment” in the future.
- Virtual machines that can simulate today’s computer systems may be the best answer for experiencing these documents in the future.
XML Myths
XML for …

- XML is **not** a markup language (**extensible** markup language)
- XML is a set of rules and tools for making markup languages
  
  - The complexity of XML is dictated by what it is used for not by XML syntax

See:  [http://xml.coverpages.org/xmlApplications.html](http://xml.coverpages.org/xmlApplications.html)
<businesscard>
    <name sex="M" suffix="Jr.">Joe Smith</name>
    <title>Senior Principal Scientist</title>
    <company>
        <name>Adobe Systems Incorporated</name>
        <address>345 Park Avenue</address>
        <address>San Jose, Ca 95110-2704</address>
        <phone>(408) 536-6000</phone>
        <fax>(408) 537-4042</fax>
    </company>
    <email>jsmith@adobe.com</email>
</businesscard>
An Element

```
<businesscard>
  <name sex="M" suffix="Jr.">Joe Smith</name>
  <title>Senior Principal Scientist</title>
  <company>
    <name>Adobe Systems Incorporated</name>
    <address>345 Park Avenue</address>
    <address>San Jose, Ca 95110-2704</address>
    <phone>(408) 536-6000</phone>
    <fax>(408) 537-4042</fax>
  </company>
  <email>jsmith@adobe.com</email>
</businesscard>
```
Element Named Attributes

Named Attributes

```xml
<businesscard>
  <name sex="M" suffix="Jr.">Joe Smith</name>
  <title>Senior Principal Scientist</title>
  <company>
    <name>Adobe Systems Incorporated</name>
    <address>345 Park Avenue</address>
    <address>San Jose, Ca 95110-2704</address>
    <phone>(408) 536-6000</phone>
    <fax>(408) 537-4042</fax>
  </company>
  <email>jsmith@adobe.com</email>
</businesscard>
```
<svg xml:space="preserve" width="612" height="792">
  <path style="fill-rule:nonzero; d="M297,264V64.5H64.5V264H297"/>
  <path style="fill-rule:nonzero;fill:#FF0000; d="M241.5,320.5c55.228,0,100-44.772,100-100c0-55.228-44.772-100-100-100c-55.228,0-100,44.772-100,100c0,55.228,44.772,100,100,100"/>
  <path style="fill-rule:nonzero;fill:#FFFF00;" d="M260.406,204.138l-97.894,175.38l232.207,0.087L260.406,204.138"/>
</svg>
<svg xml:space="preserve" width="612" height="792">
    <path style="fill-rule:nonzero;"
        d="M297,264V64.5H64.5V264H297"/>
    <path style="fill-rule:nonzero;fill:#FF0000;"
        d="M241.5,320.5c55.228,0,100-44.772,100-100c0-55.228-44.772-100-100-100c-55.228,0-100,44.772-100,100c0,55.228,44.772,100,100,100"/>
    <path style="fill-rule:nonzero;fill:#FFFF00;" d="M260.406,204.138l-97.894,175.38l232.207,0.087L260.406,204.138"/>
</svg>

Human Readable?
<xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl">
  <xsl:template match="/">
    <xsl:for-each select="./body/p">
      <p>
        <style>bchar { font-size:"xxlarge";}</style>
        <xsl:value-of select="."/>
      </p>
    </xsl:for-each>
  </xsl:template>
</xsl:stylesheet>
XML for Documents

- Example: DocBook
  - An extensive document element set:
    - Abstract, preface, chapter, paragraphs, quotes, lists (5 kinds), figures, headings, footnotes, examples, bibliography, appendix, colophon, cross referencing, media objects, etc.


- But this does not provide:
  - Birth certificate
  - Invoice
  - Legislative acts

- Each needs its own XML for …
XML Documents

- But what about OOXML and ODF
  - XML documents?

- These are not single XML files but ZIP archive packages
  - Main textual content are XML subfiles
  - Resources: Fonts, Images, Figures

- Allows binary data to remain in proper form
- Provides compression for XML subfiles (Flate)
- Random access at subfile granularity

- The XML design point is marked up text
Document Format Complexity

- What will matter in 50 years?
  - XML is human readable
  - Standards describing the formats are preserved
    - Can implement readers
  - Ability to read the electronic representation
    - Media
    - Format
  - Continuous use during the 50 years
Conclusions
Conclusions

• First determine the planned future use for the archived documents
  • Editable requirement primary: use ODF or OOXML or DocBook
  • Appearance preservation important: use PDF/A

• Be more refined when judging XML

• Consider XML for what
  • XML is not compact
  • XML is not good for binary data
  • Consider one of the ZIP packaged solutions

• Many other things beside file format need be considered
Thank You for Your Attention!

Any questions?

Jim King’s Web pages:
http://www.adobe.com/technology/people/sanjose/king.html
http://blogs.adobe.com/insidepdf

Leonard Rosenthal’s Blog:
http://www.acrobatusers.com/blogs/leonardr/

PDF Tutorial:

and PDF imaging model:

Contact us via: – mailto: jking@adobe.com