Document Metadata:
document technical metadata for digital preservation

By
Carol Chou - Florida Digital Archive (FDA)
Andrea Goethals - Harvard Library (HL)

March 18, 2009
Rev. November 30, 2012

Table of Contents

Change History

Date	Person	Changes
November 28, 2012	Andrea Goethals	• Updated introduction • Added a change history • Added new elements: references, documentMetadataExtension • Added new sub-elements to the Features element: hasFixedLayout, has Audio, hasVideo, hasScripts • Switched out the schema to the schema that was revised on October 29, 2012 and November 16, 2012
November 30, 2012	Carol Chou	• Update introduction.

Introduction

The Florida Virtual Campus (FLVC)[1] provides digital preservation for the eleven public universities in Florida. Since it was established in 2005, FDA has ingested over forty-two millions files with over one hundred and nineteen terabytes of data. There are approximately 250,000 files in document formats such as PDF, Word and OpenDocument Text Format. Most of these documents come from the Electronic Thesis and Dissertation (ETD) collections in those universities. Ensuring all FDA collections remain usable and renderable is one of the critical missions for FDA.

Extracting technical metadata from documents is essential as it can aid in characterizing the kinds of documents in our preservation collections; listing document properties that may hinder preservation (encryption, external fonts, etc); and providing requirements in selecting tools/facilities for document transformation including normalization and migration. In addition, Document technical metadata can be used to verify the result of document transformations, ensuring the properties of the original document are preserved and properly transformed to the new document format.

There are currently many metadata standards for various format groups. For images, there is MIX (NISO Metadata for Images in XML[2]). For text-based formats such as plain-text, XML, DTD and HTML, there is the TextMD (Technical Metadata for Text[3]) schema. In terms of audio, there is the AES57-2011[4] (AES Audio Object) schema from the Audio Engineering Society, and the AMD schema[5] by the Library of Congress. When it comes to document formats such as PDF, Word or OpenDocument Text, it came to our attention that there wasn't any technical metadata standard to follow.

Though JHOVE provides a metadata extraction function for PDF and presents the extracted PDF metadata in JHOVE schema, the extracted metadata is massive and is expressed in a page-by-page manner. This document metadata schema is intended to be simpler and may be applied to document formats other than PDF. The document metadata schema may be expressed in XML or database form. We will also develop a style-sheet to convert PDF metadata in the JHOVE schema into the document metadata schema.

Andrea Goethals at Harvard Library (HL) also expressed the need for a document metadata schema. HL enhanced their preservation repository, the Digital Repository System (DRS), to accept born-digital content, including documents, and saw this document metadata schema as the first step towards preserving documents for the long-term. Together, we developed this document metadata schema for the use of FDA and the DRS. We hope to gather input from

1 In July 2012 the Florida Center for Library Automation (FCLA) combined with 3 other organizations to become the Florida Virtual Campus (FLVC).

2 See http://www.loc.gov/standards/mix/

3 See http://www.loc.gov/standards/textMD/

4 See http://www.aes.org/publications/standards/search.cfm?docID=84

5 See http://www.loc.gov/rr/mopic/avprot/metsmenu2.html

the preservation community to enhance the document metadata schema which may also be useful to other trusted repositories as well.

Applicable Formats

This metadata schema was developed based on a small set of popular document formats but should be generally applicable to formats that:

- are primarily text.
- allow creators a choice of fonts, colors, text size, backgrounds.
- support embedded multimedia (images, sounds, video etc).
- may contain application specific features.
- may be intended for typeset documents.

These formats include but are not limited to:

Format	MIME Type	File Extension	Native Applications
Microsoft Word	application/msword application/ms-word (This second MIME type isn't correct but is sometimes used)	doc	Microsoft Word and Microsoft Office Word
Portable Document Format	application/pdf	pdf	Adobe Writer
OpenDocument Text	application/vnd.oasis.opendocument.text	odt	OpenOffice.org2.0 / StarOffice 8 and later
Writer 6.0 Document	application/vnd.sun.xml.writer	sxw	OpenOffice.org1.0 / StarOffice6.0 and later
StarWriter 5.x Document	application/vnd.stardivision.writer	sdw	StarOffice 5.x
StarWriter 4.x Document	application/x-starwriter	sdw	StarOffice 4.x
WordPerfect Document	application/vnd-wordperfect	wpd	WordPerfect and WordPerfect Office
Works Text Document	application/vnd.ms-works	wps	Microsoft Works

For each metadata element listed in the data dictionary, the document formats are listed that are known to contain either the associated metadata values directly in the file or that could be determined indirectly by parsing the files.

Data Dictionary

The data dictionary describes the semantic meaning and constraints of document-specific metadata. The document specific metadata are those document properties that are deemed preservation-worthy and pertain to most document formats. Some elements are included because they will aid in evaluating the completeness of the content after transformations (e.g. number of pages). Other elements are included because they will aid in selecting or aggregating documents for risk analysis, preservation or delivery planning (e.g. Features).

Please note that general preservation and descriptive metadata may also be extracted from documents. This metadata includes size, encryption, title, author/creator, create-date, copyright, digital signature, protection/permission, etc. and may be recorded by using standard preservation schema such as PREMIS and MODS.

Semantic unit	PageCount
Semantic components	None
Description	Total number of pages in the document
Data Constraint	Min 1
Obligation	Mandatory
Cardinality	1
Characteristic	Structure
Note	

Semantic unit	WordCount
Semantic components	None
Description	Total number of words in the document
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	This element is included in this schema because it can be valuable for evaluating the completeness of the content after transformations. Caution must be used with this element however because tools and applications that can determine the number of words in a document do not always use the same algorithm for determining this value.

Semantic unit	CharacterCount
Semantic components	None
Description	Total number of characters in the document
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	This element is included in this schema because it can be valuable for evaluating the completeness of the content after transformations. Caution must be used with this element however because tools and applications that can determine the number of characters in a document do not always use the same algorithm for determining this value.

Semantic unit	ParagraphCount
Semantic components	None
Description	Total number of paragraphs in the document
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	This element is included in this schema because it can be valuable for evaluating the completeness of the content after transformations. Caution must be used with this element however because tools and applications that can determine the number of paragraphs in a document do not always use the same algorithm for determining this value.

Semantic unit	LineCount
Semantic components	None
Description	Total number of lines in the document
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	This element is included in this schema because it can be valuable for evaluating the completeness of the content after transformations. Caution must be used with this element however because tools and applications that can determine the number of lines in a document do not always use the same algorithm for determining this value.

Semantic unit	TableCount
Semantic components	None
Description	Total number of tables in the document
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	This element is included in this schema because it can be valuable for evaluating the completeness of the content after transformations. Caution must be used with this element however because tools and applications that can determine the number of tables in a document do not always use the same algorithm for determining this value.

Semantic unit	GraphicsCount
Semantic components	None
Description	Total number of graphics (* andrea) in the document
Data Constraint	Min 0
Obligation	Optional
Cardinality	1
Characteristic	Structure
Note	This element is included in this schema because it can be valuable for evaluating the completeness of the content after transformations. Caution must be used with this element however because tools and applications that can determine the number of graphics in a document do not always use the same algorithm for determining this value.

Semantic unit	Language
Semantic components	None
Description	A language identifier specifying the natural language used in the document
Data Constraint	String (or some kind of controlled vocabulary like ISO 639-2 alpha-3 language codes)
Obligation	Optional
Cardinality	0 - N
Characteristic	Content
Note	

Semantic unit	Font
Semantic components	FontName isEmbedded
Description	A list of fonts used in the document
Data Constraint	Container
Obligation	Mandatory
Cardinality	1 - N
Characteristic	Content, Appearance
Note	This element allows a repository to store the names of all fonts used in a document. Some repositories may choose to store only the non-embedded fonts. The use of non-embedded fonts may hinder the long term preservation of the documents. For example, a document encoded with a proprietary non-embedded math font may not be migrated due to unavailability of the specific math font. It is recommended that repositories record at least the non-embedded fonts to assist in identifying the documents with potential long-term preservation risks.

Semantic unit	FontName
Semantic components	None
Description	Name of a font
Data Constraint	String
Obligation	Optional
Cardinality	1
Characteristic	Content, Appearance
Note	

Semantic unit	IsEmbedded
Semantic components	None
Description	An indication of whether or not a font is embedded in a document.
Data Constraint	Y, N
Obligation	Optional
Cardinality	1
Characteristic	Content, Appearance
Note	

Semantic unit	Reference
Semantic components	None
Description	An URI referenced within the document
Data Constraint	String
Obligation	Optional
Cardinality	0 - N
Characteristic	Content
Note	This element makes it possible for a repository to archive the de-referenced URIs to have more of the context of the document.

Semantic unit	Features
Semantic components	None
Description	Additional document features
Data Constraint	isTagged, hasOutline, hasThumbnails, hasLayers, hasForms, has Annotations, hasAttachments, hasTransparancy, hasFixedLayout, has Audio, hasVideo, hasScripts
Obligation	Optional
Cardinality	0 - N
Characteristic	isTagged: structure hasOutline: behavior, appearance hasThumbnails: appearance hasLayers: appearance hasForms: content hasAnnotations: content hasAttachments: structure, behavior hasTransparency: appearance hasFixedLayout: appearance has Audio: content hasVideo: content hasScripts: appearance, behavior
Note	

Semantic unit	documentMetadataExtension
Semantic components	Defined externally
Description	A container to include semantic units defined outside of DocumentMD
Data Constraint	Container
Obligation	Optional
Cardinality	0 - N
Characteristic	
Note	This element makes it possible for a repository to include externally defined semantic units.

Document Metadata schema

```
<!--

        Editor: Florida Center for Library Automation (FCLA) and Harvard University Library (HUL)
        Released: March 17, 2009
        June 2, 2010 Change: Make PageCount element optional to account for PDFs whose pageCount cannot
be determined.
        Oct 29, 2012 Change: Per request from National Library of France,
                1. Add optional Reference element for URLs declared in the document.
                2. Add hasFixedLayout, hasAudio, hasVideo and hasScripts features.
                3. Add optional documentMetadataExtension section for extending DocumentMD.
        November 16, 2012 Change: Make documentMetadataExtension repeatable
-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:docmd="http://www.fcla.edu/dls/md/docmd"t
arget-
Namespace="http://www.fcla.edu/docmd" elementFormDefault="qualified" attributeFormDefault="unqualified">
    <xs:element name="document">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="PageCount" minOccurs="0" maxOccurs="1">
                    <xs:simpleType>
                        <xs:restriction base="xs:integer">
                            <xs:minInclusive value="0"/>
                        </xs:restriction>
                    </xs:simpleType>
                </xs:element>
                <xs:element name="WordCount" minOccurs="0" maxOccurs="1">
                    <xs:simpleType>
                        <xs:restriction base="xs:integer">
                            <xs:minInclusive value="0"/>
                        </xs:restriction>
                    </xs:simpleType>
                </xs:element>
                <xs:element name="CharacterCount" minOccurs="0" maxOccurs="1">
                    <xs:simpleType>
                        <xs:restriction base="xs:integer">
                            <xs:minInclusive value="0"/>
                        </xs:restriction>
                    </xs:simpleType>
                </xs:element>
                <xs:element name="ParagraphCount" minOccurs="0" maxOccurs="1">
                    <xs:simpleType>
                        <xs:restriction base="xs:integer">
                            <xs:minInclusive value="0"/>
                        </xs:restriction>
                    </xs:simpleType>
                </xs:element>
                <xs:element name="LineCount" minOccurs="0" maxOccurs="1">
                    <xs:simpleType>
                        <xs:restriction base="xs:integer">
                            <xs:minInclusive value="0"/>
```

```
                </xs:restriction>
            </xs:simpleType>
        </xs:element>
        <xs:element name="TableCount" minOccurs="0" maxOccurs="1">
            <xs:simpleType>
                <xs:restriction base="xs:integer">
                    <xs:minInclusive value="0"/>
                </xs:restriction>
            </xs:simpleType>
        </xs:element>
        <xs:element name="GraphicsCount" minOccurs="0" maxOccurs="1">
            <xs:simpleType>
                <xs:restriction base="xs:integer">
                    <xs:minInclusive value="0"/>
                </xs:restriction>
            </xs:simpleType>
        </xs:element>
        <xs:element name="Language" minOccurs="0" maxOccurs="unbounded" type="xs:string"/>
        <xs:element name="Font" minOccurs="0" maxOccurs="unbounded">
            <xs:complexType>
                <xs:attribute name="FontName" type="xs:string"/>
                <xs:attribute name="isEmbedded" type="xs:boolean"/>
            </xs:complexType>
        </xs:element>
        <xs:element name="Reference" minOccurs="0" maxOccurs="unbounded" type="xs:string"/>
        <xs:element name="Features" minOccurs="0" maxOccurs="unbounded">
            <xs:simpleType>
                <xs:restriction base="xs:string">
                    <xs:enumeration value="isTagged"/>
                    <xs:enumeration value="hasOutline"/>
                    <xs:enumeration value="hasThumbnails"/>
                    <xs:enumeration value="hasLayers"/>
                    <xs:enumeration value="hasForms"/>
                    <xs:enumeration value="hasAnnotations"/>
                    <xs:enumeration value="hasAttachments"/>
                    <xs:enumeration value="useTransparency"/>
                    <xs:enumeration value="hasFixedLayout"/>
                    <xs:enumeration value="hasAudio"/>
                    <xs:enumeration value="hasVideo"/>
                    <xs:enumeration value="hasScripts"/>
                </xs:restriction>
            </xs:simpleType>
        </xs:element>
        <xs:element name="documentMetadataExtension" minOccurs="0" maxOccurs="unbounded">
            <xs:complexType>
                <xs:sequence>
                    <xs:any namespace="##any" processContents="lax" minOccurs="0"
                    maxOccurs="unbounded"/>
                </xs:sequence>
            </xs:complexType>
        </xs:element>
    </xs:sequence>
</xs:complexType>
```

```
    </xs:element>
</xs:schema>
```

Appendix A: A sample PREMIS document with embedded docmd schema

```
<premis xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="info:lc/xmlns/premis-v2"
xsi:schemaLocation="info:lc/xmlns/premis-v2 http://www.loc.gov/standards/premis/draft-schemas-2-0/premis-v2-
0.xsd" version="2.0">
    <object xsi:type="file">
        <objectIdentifier>
            <objectIdentifierType>DAITSS2</objectIdentifierType>
        <objectIdentifierValue>/Users/Carol/Desktop/work/testdata/pdf/etd.pdf</objectIdentifierValue>
        </objectIdentifier>
        <objectCharacteristics>
            <compositionLevel>0</compositionLevel>
            <size>3312509</size>
            <format>
                <formatDesignation>
                    <formatName>PDF</formatName>
                    <formatVersion>1.3</formatVersion>
                </formatDesignation>
                <formatRegistry>
                    <formatRegistryName>PRONOM</formatRegistryName>
                    <formatRegistryKey>fmt/17</formatRegistryKey>
                </formatRegistry>
            </format>
            <creatingApplication>
                <creatingApplicationName>THESIS (Electronic thesis).doc - Microsoft
Word</creatingApplicationName>
                <dateCreatedByApplication>Tue Apr 24 16:22:40 EDT 2001</dateCreatedByApplication>
            </creatingApplication>
            <fixity>
                <messageDigestAlgorithm>MD5</messageDigestAlgorithm>
                <messageDigest>6bf12f206a3e70c88cfe2aa5213dd227</messageDigest>
            </fixity>
            <objectCharacteristicsExtension>
                <doc xmlns="http://www.fcla.edu/dls/md/docmd.xsd">
                    <document>
                        <PageCount>123</PageCount>
                        <Font FontName="Arial" isEmbedded="false"/>
                        <Font FontName="TimesNewRoman,BoldItalic" isEmbedded="false"/>
                        <Font FontName="BookmanOldStyle" isEmbedded="false"/>
                        <Font FontName="Arial,Bold" isEmbedded="false"/>
                        <Font FontName="TimesNewRoman,Italic" isEmbedded="false"/>
                        <Feature>hasThumbnails</Feature>
                    </document>
                </doc>
            </objectCharacteristicsExtension>
        </objectCharacteristics>
    </object>
</premis>
```