Adobe's Extensible Metadata Platform (XMP): Background

[DRAFT -- Caroline Arms, 2011-11-30]

Contents

Introduction

Adobe's XMP (Extensible Metadata Platform: http://www.adobe.com/products/xmp/) is a mechanism for embedding metadata into content files. For example. an XMP "packet" can be embedded in PDF documents, in HTML and in image files such as TIFF and JPEG2000 as well as Adobe's own PSD format native to Photoshop. In September 2011, XMP was approved as an ISO standard.[ISO 16684-1: Graphic technology -- Extensible metadata platform (XMP) specification -- Part 1: Data model, serialization and core properties]

XMP is an application of the XML-based Resource Description Framework (RDF; http://www.w3.org/TR/2004/REC-rdf-primer-20040210/), which is a generic way to encode metadata from any scheme. RDF is designed for it to be easy to use elements from any namespace.

An important application area is in publication workflows, particularly to support submission of pictures and advertisements for inclusion in publications. The use of RDF allows elements from different schemes (e.g., EXIF and IPTC for photographs) to be held in a common framework during processing workflows.

There are **two ways to get XMP metadata into PDF documents**:

- **manually** via a customized File Info panel (or equivalent for products from vendors other than Adobe). This approach would be most appropriate for individually created PDFs at the time of creation. A Custom Info Panel can be basic or highly functional. A basic panel may require users entering data to follow rules very carefully and type accurately; elements with a few possible values can be entered via drop-down menus. A highly functional panel can be constructed (by someone with programming skills) to support lookup from remote controlled vocabularies and perform element-specific validation for syntax (for example for dates). For more on the creation of Custom Info Panels, see Appendix C.
  Once a Custom Info Panel has been designed, adding it to Adobe products in use at LC to create PDFs, such as Acrobat Pro, InDesign, Photoshop, is straightforward and can be done without re-installation of the applications.

- by **preparing a chunk of metadata with the appropriate markup in a file** and using a tool that embeds that chunk in an existing PDF. This approach is more compatible with batch operations, whether as part of an automated workflow or for retrospective upgrade.

Important: XMP metadata in a PDF is not visible to users of Adobe Reader. Metadata that LC wishes to have visible to users of PDFs that it creates must be in the four Basic PDF metadata fields and/or presented within the document itself.

There is limited adoption of XMP among communities with which the Library of Congress interacts. The digital library community is not currently making use of XMP metadata in PDFs and has not developed tools either for embedding or for indexing XMP. For still images, both the archival community and the professional photography community are developing tools and encouraging the use of XMP as part of guidelines for best practice. In the scholarly publishing community a few publishers have experimented with embedding XMP metadata and adoption may increase if publishers embrace the new CrossMark service from CrossRef [http://www.crossref.org/crossmark/index.html].

Adobe's XMP toolkits

Adobe makes available a free Software Development Toolkit (SDK) under the BSD license. The SDK consists of two libraries. The XMPCore library supplies an API for parsing, manipulating, and serializing metadata according to the XMP data model. The XMPFiles library contains a number of file handlers that know how to efficiently access the XMP in almost 30 specific file formats, including PDF. Both libraries are available in C++ for Windows, Mac OS, and Linux. XMPCore is also available in a Java implementation.

Adobe also makes available a toolkit for including a custom "File Info" panel in Adobe applications such as Acrobat Pro. The Custom File Info SDK provides documentation and samples on how to create a user interface panel for viewing and editing custom metadata.
Examples:

- A simple example of use of this feature is the File Info Panel for a Creative Commons license. See http://wiki.creativecommons.org/Adobe_Metadata_Panel#Adobe_CS4_and_CS5 and http://johnbishopimages.com/?xmp
- A more complex example, the Information Interchange Model (IIM) from the International Press Telecommunications Council (IPTC) is illustrated at http://www.poundhillsoftware.net/Schema_IPTC_IIM_for_XMP.htm

Links to Adobe pages about XMP adoption

- Industry standards groups powered by XMP.
  http://www.adobe.com/products/xmp/standards.html
  AdsML; Creative Commons; Digital Image Submission Criteria (DISC); Dublin Core Metadata Initiative (DCMI); IPTC; Metadata Working Group (Canon, Microsoft,

Sony, Nokia, etc., focused on still images); Picture Licensing Universal System (PLUS) Coalition; Publishing Requirements for Industry Standard Metadata (PRISM)

- Partners. http://www.adobe.com/products/xmp/partners.html
  Mainly software vendors providing systems and services for publishing and media industries.

Appendix A

Mapping of PDF Document Info (basic metadata) to XMP properties

The following table is extracted from the Part 3 of the XMP specification: Storage in Files (version of July 2010, retrieved from Adobe website September 2011). Notice that use of dc:creator , dc:title, dc:subject, and dc:description elements is constrained by the automatic mapping between them and the PDF elements that happens within Adobe applications. The automatic mapping for dc:title and for dc:description is probably acceptable. However, the operational mapping constraints for dc:creator and dc:subject may create sufficient problems that it would be advisable not to use these elements, but to use equivalent element from a different namespace (possible an LC-specific one).

Table 18 — Mapping of PDF keys to XMP properties

PDF Document Info key	XMP metadata property	Mapping notes
Title	**dc:title**	The **Title** key maps to the first of the alternatives given in the **dc:title** property.
Author	**dc:creator**	The **Author** key maps to the first of the creators listed in the **dc:creator** field. Alternatively (available by user action in the Acrobat 7 UI), maps to a concatenated list of the creators listed in the **dc:creator** field separated by a standard separator character such as semicolon.
Subject	**dc:description**	The **Subject** key maps to the first of the alternatives given in the **dc:description** property.
Keywords	**pdf:Keywords**	The XMP properties **dc:subject** and **pdf:Keywords** have historically been separate. In Acrobat 7, Adobe allows user intervention to set them to corresponding values, where the value in the

		PDF schema (and in the DocInfo) is set to a delimiter-separated concatenation of the bag of values found in the dc:subject value.
Creator	**xmp:CreatorTool**	
Producer	**pdf:Producer**	
CreationDate **ModDate**	**xmp:CreateDate** **xmp:ModifyDate**	Info dictionary dates are in ISO/IEC 8824 format, XMP dates are in ISO 8601 format.
Trapped	**pdf:Trapped**	The **Trapped** key is Boolean.

Note: Trapped is an important element for the submission of graphic materials to magazines.

Appendix B

Software applications that can read or write XMP metadata in PDFs

XMP is increasingly supported in still image tools, including open source software. Support for working with XMP in PDFs is primarily provided through commercial software aimed at the publishing industry, where it supports workflow for submitting content and advertisements to publishers.

Adobe Products

All products in the Creative Suite family, including Acrobat Pro (not Acrobat Standard), InDesign, Photoshop, etc. can embed XMP metadata via Info Panels. Although these products are in use at LC, no-one has used them for embedding XMP metadata.

Each version of the Creative Suite comes with a new/increased selection of Info Panels from other entities that have generated market demand. Custom Info Panels can be developed; see Appendix C.

Other Commercial Products

- **LuraDocument PDF Compressor**, from LuraTech, (one of the products in use at LC for creating PDFs from scanned images) can embed XMP metadata in a PDF if the metadata chunk is prepared outside the product and stored in a file named appropriately for automatic embedding if the option to embed is chosen and the file is present.
  See http://www.luratech.com/home/products/software-and-solutions-for-document-processing/document-and-data-conversion-software/luradocument-pdf-compressor.html
- **MetaGrove applications, from Poundhill Software**. MetaGrove Developer can be used to build a custom XMP schema and Custom Info Panels for Adobe applications. MetaGrove plug-ins exist for different Adobe applications, such as Photoshop. InDesign, and Acrobat.
  See http://www.poundhillsoftware.net/dev.htm

Open Source toolkits and applications

- **ExifTool** (command-line application written in Perl)
  See http://owl.phy.queensu.ca/~phil/exiftool/struct.html

  This product is designed to read files, manipulate metadata, and write new files in the same format. Efficient batch operations. Already in use at LC, e.g. for preparations of PDFs for the World Digital Library.
- **iText** (SDK, available in Java and C#)
  See http://itextpdf.com/book/chapter.php?id=12

  This product is aimed at generating PDF files from text sources, including embedding metadata (Basic PDF and XMP). Used by PdfLicenseManager application. Also

includes XMP extraction capability. Already used at LC by Michael Ferrando (ITS) for embedding Basic PDF metadata into PDF files.

- **Python XMP Toolkit** (C/C++ library based on Adobe's XMP Toolkit, distributed under BSD license)
  See http://www.spacetelescope.org/static/projects/python-xmp-toolkit/docs/index.html and http://code.google.com/p/python-xmp-toolkit/

  Developers (from European participants in the Hubble Space Telescope team) use Mac OS X and Linux. Not tested on Windows.

- **Apache Tika** (SDK, available in Java)
  The Apache Tika™ toolkit detects and extracts metadata and structured text content from various document types (including PDF) using existing parser libraries.

  See http://tika.apache.org/

- **PDFmark**. (Java-based command-line application)
  Experimental open source tool designed for adding CrossRef metadata to a PDF. Includes ability to add XMP metadata to a PDF by passing the tool a pre-generated XMP file. Hence, it may be usable for non-CrossRef metadata.

  See http://labs.crossref.org/pdfmark/pdfmark.html

- **PDF Information Editor** and **Advanced PDF Tools** from verypdf.com (Applications for Windows)
  Includes ability to add XMP metadata to a PDF by passing the tool a pre-generated XMP file.

  See http://www.verypdf.com/pdfinfoeditor/index.html

Free application for viewing or extracting XMP metadata in PDFs

- **XMPexplorer** (for Windows)
  From http://www.logicmighty.co.uk/
- **Embedded Metadata Explorer** (web-based)
  http://embedmydata.com/

Appendix C

Creating Custom Info Panels for embedding XMP metadata

Facts about Custom Info Panels for Adobe applications

- Custom Info Panels can be developed to work with all Creative Suite (CS) tools (including Photoshop, InDesign, Adobe Bridge, and Acrobat Pro). As of September LC installs tools from CS4 when requested. Current version from Adobe is CS5.
- Custom Info Panels for Acrobat Pro must use an old mechanism (and form of files for the panels) used for old (CS2 and CS3) versions of the other Adobe tools. Hence LC would need to create two forms of panels. Installation of Custom Info Panels so that they can be used in Acrobat Pro 9 in LC-configured computers running Windows XP has been successfully tested for feasibility.
- Adobe Bridge can be used to create and apply metadata through templates to collections of PDF documents (as well as image files). Templates can use fields from Custom Info Panels.
- Basic panels can be developed using a text editor. This process requires some skills typical of a programmer (getting the configuration syntax precise) but can be performed by copying and adapting examples.
- More complex panels for applications other than Acrobat Pro can include support for searching controlled vocabularies in order to populate metadata elements and element-specific syntax validation. These require use of Javascript or Flash.

Issues for experimentation or exploration

- How best to support repeatable fields.
- How to allow commas within values. This is necessary to support multiple authors/creators/contributors and names used as subjects.
- To what extent is functionality available via Acrobat Pro less than that available via other Adobe tools.
- Can sub-elements of structured values (such as hierarchical geographic terms) be represented as such or is a string with specific separators (such as "--") the most convenient approach?
- Are basic panels adequate or would acquisition of a specialized tool for developing panels or use of a consultant be worthwhile?

Source of experience in building Custom Info Panels

- Greg Reser from UCSD has developed panels intended for embedding VRA Core metadata in image files. He used the Adobe Flash Builder product to build panels for the CS4/5 applications (i.e. not for Acrobat Pro). He would be happy to demonstrate what he has done and discuss dos and don'ts. He could also recommend a contractor for building a panel that included vocabulary lookup and validation for individual elements.

Alternative approach to consider

- Rather than using Custom Info Panels, might it be more cost-effective to build a data entry tool independent of Adobe applications (perhaps using Adobe Air, Flash Builder or JavaScript) that could be used to enter metadata for PDFs and for HTML pages? Data entry would use common forms, and take advantage of vocabulary lookup and element-specific syntax validation. Export would be possible in the correct markup for XMP and to use as meta tags. There are several tools available that claim to take a prepared chunk of XMP and add it to a PDF.