

OIS Recommendations for PDF Files Created for Long-term Preservation and Access

Andrea Goethals, Harvard University Library

Last modified: 02/11/2010

Table of Contents

1 Change History.....	1
2 PDF Formats accepted by the DRS.....	1
3 PDF Best Practices.....	2
General Practices.....	2
PDF Feature Recommendations.....	2
Very important practices.....	2
Good practices.....	3

1 **Change History**

- July 30, 2009 Initial draft
- Feb. 11, 2010 Added a change history section and a caveat in section 2 about encrypted PDFs

2 **PDF Formats accepted by the DRS**

The DRS accepts files in any variation of the PDF format. However, to facilitate the long-term preservation, accessibility and rendering of PDF files, OIS recommends that PDF files be created in or converted to the most preferred format possible:

Most Preferred	PDF/A-1a
·	PDF/A-1b
·	Any valid PDF format that conforms with the recommendations found in the <i>PDF Feature Recommendations</i> section of this document.
·	Any valid PDF format
·	Invalid PDF
Least Preferred	

Although the DRS accepts any variation of the PDF format, the DRS will not accept PDF documents with embedded security measures that hamper future use or the copying and transformations needed for preservation.

3 **PDF Best Practices**

General Practices

- Validate your PDFs
 - Run a representative set of your PDF files against multiple PDF validation tools. In testing OIS has found that the Preflight tool within Adobe Acrobat Professional 9 is able to find PDF errors not found by Adobe Acrobat Professional 8 nor JHOVE 1.1¹. PDF validation is especially important when developing the creation work flow for your PDF files so that you can verify that the tools and process you are using will result in preservable PDFs.
- Improve your PDFs
 - Use a tool on your PDF files to convert them to a PDF/A format (if possible), fix any problems with them, add metadata to the files, etc. before depositing them to the DRS. Any conversions should replicate the exact content and quality of the source document in the new document -- if possible test this with a representative set before performing batch conversions.

PDF Feature Recommendations

Very important practices

- Remove any security restrictions (password, certificate, etc.) on the documents. These restrictions could hamper future use or transformations of the documents, and can interfere with screen readers. This is also a PDF/A requirement.
- Embed and subset *all* document fonts, including the 14 standard Type 1 fonts (Times-Roman, Times-Bold, Times-italic, Times-BoldItalic, Symbol, Helvetica, Helvetica-Bold, Helvetica-Oblique, Helvetica-BoldOblique, Courier, Courier-CBold, Courier-Oblique, Courier-BoldOblique, ZapfDingbats). This is also a PDF/A requirement.
- If possible, create tagged PDFs. Tags describe logical structural aspects (paragraphs, lists, tables, links, illustrations, etc.) and are used by rendering applications and screen-reader devices. This is also a PDF/A-1a but not a PDF/A-1b requirement.
 - Provide alternative text descriptions for all non-text document elements, e.g. images, that are part of the document's content. You do not need to provide alternative text descriptions for images that provide decorative elements to the document.
 - Specify the natural language primarily used in the document. You can set the language for a document in Acrobat Pro through File->Properties, Advanced Tab, then select a Language from the pull-down list in the Reading Options section. For more information see Adobe's "Set the Document Language" page at http://help.adobe.com/en_US/Acrobat/9.0/Professional/WS58a04a822e3e50102bd615109794195ff-7cee.w.html
- For PDFs created from scanned text, include the OCR text in the file so that the text can be searched, copied and read by screen readers and in search interfaces.

1 Since this was written there have been improvements to later versions of JHOVE to better recognize PDF/A.

Good practices

- Avoid lossy compression algorithms. This is a recommendation but not a requirement for PDF/A.
 - The lossy compression algorithms or filters include DCTDecode (JPEG) and potentially JBIG2Decode and JPXDecode (JPEG2000) (these two can be lossy or lossless).
- If it is important that the images in the PDF be high resolution, do not downsample (decrease the number of pixels in) the images. Downsampling is an option in PDF optimizing but can lead to poor quality images. This is a recommendation but not a requirement for PDF/A.
- If possible, do not use transparency in images. This is a PDF/A requirement.
- Do not use fonts that can't be legally embedded.
- For any hyperlinks in the document, write out the URIs, for example: AIIM International (<http://www.aiim.org>) instead of [AIIM International](#).
- Embed metadata in your PDF documents. Some metadata will be automatically added for you to the document by your PDF creation software. You can add additional metadata manually per PDF or as part of a batch process. The following metadata is recommended:
 - Title - either pdf:Title and/or dc:title and/or xmp:Title *
 - Author - either pdf:Author and/or dc:creator and/or xmp:Author *
 - Subject - either pdf:Subject and/or dc:description and/or xmp:Description *
 - Keywords - pdf:Keywords and/or dc:subject and/or xmp:Keywords *
 - Creator - either pdf:Creator and/or xmp:CreatorTool *
 - Producer - pdf:Producer
 - CreationDate - either pdf:CreationDate and/or xmp:CreateDate *
 - ModDate - either pdf:ModDate and/or xmp:ModifyDate *

* When equivalent multiple metadata elements are present in a document (e.g. pdf:Title and dc:title), their values should be equivalent.

For DRS users: The PDF/A specification requires use of XMP metadata, but DRS tools will recognize and extract XMP as well as the non-XMP metadata listed above.

- Do not use the Launch, Sound, Movie, ResetForm, ImportData, or JavaScript actions. This is also a PDF/A requirement.
- Do not embed nor attach files to PDF documents.