Development of archival stream standards for born-digital documents
(Final Recommendation)

Overview

To date, the majority of objects in the Digital Library Repository/RUcore consist primarily of data streams obtained through the digitization of physical, or analog, objects. Development of preservation standards for such objects proved to be relatively simple, as the imaging industry laid much of the groundwork for us in terms of standardization across platforms. Further, development of future standards for digitized images, sound and video continues in an organized and orderly fashion, giving us plenty of time to contemplate migration to newer and better preservation formats.

Unfortunately, the same cannot be said for “born-digital” documents – that is to say, objects which originated in digital form. File formats for such objects vary widely, and the challenge is upon us to identify a uniform set of file formats that we can adopt for preservation purposes.

Proposed Preservation Format Strategy: Multiple standards in play

At present, there are a number of formats developed by various consortia that attempt to solve the problem of maintaining a persistent document standard. Our choice of standards must hinge on its ability to endure as technological advances continue to develop, and a widespread acceptance is key to ensuring easy migrating to newer standards when the time comes to retire existing choices.

That said, our timing timetable for developing a standard is inopportune, in that a standards war is brewing as document format evolve. Currently, there is competition between the open source community and commercial vendors to entice organizations to adopt two competing document formats. While the two formats each have a broad support base and share similar characteristics, they are nt interchangeable. This raises a serious concern, in that should we select one document format as our preservation standard over the other, we may find ourselves on the “losing” side of this standards war.

It is thus advisable that we consider more than one preservation datastream for born-digital objects. This strategy will permit us to build redundancy into our repository, and ensure that regardless of whether one standard “wins out” over the other, our objects will remain with at least one relevant archival datastream.

To further buttress object stability, it is proposed that the streams be augmented with a third, well-established file format, to solidify our ability to make these documents widely accessible.

The archival datastreams would consist of different file formats:

- **DS1.0 – Native Format:** The original object in its native file format, if that format is not one of the standards listed below.
- **DS2.0 - OASIS OpenDocument (ODF):** An existing, open standard for file formats in use primarily in open source an “non-Microsoft” environments
- **DS3.0 - OpenXML:** a forthcoming standard endorsed by Microsoft and a consortium of other manufacturers, intended to be widely used in the next derivative of MS Office
- **DS4.0 - Portable Document Format (PDF):** a proprietary but well-established standard, a subset of which is being considered for an open source archival document standard.

Format Analysis

OpenDocument Format: An open source and established document format

The **OpenDocument** format (**ODF**), short for the **OASIS Open Document Format for Office Applications**, is an open document file format for saving and exchanging editable office documents such as text documents (including memos, reports, and books), spreadsheets, charts, and presentations. This standard was developed by the OASIS industry consortium, based upon an XML-centric file format originally created by OpenOffice.org.

The standard was publicly developed by a variety of organizations, is publicly accessible, and can be implemented by anyone without restriction. The stated intention of the OpenDocument format is to provide an open alternative to proprietary document formats including the popular but undocumented DOC, XLS, and PPT formats used by current versions of Microsoft Office, generally regarded as the most popular document creation suite.

OpenDocument is the presently only standard for editable office documents that is currently shipping and has been vetted by an independent recognized standards body, has been implemented by multiple vendors, and can be implemented by any supplier (including proprietary software vendors as well as developers using open source software licenses such as the GNU LGPL or GNU GPL).

In addition, The American Library Association (ALA) and at least four other library organizations in the US have written a letter of support for ODF. And the state government of Massachusetts has formally adopted ODF as the standard for government-originated documents going forward.

Capabilities:

ODF is expected to handle most of the formatting challenges which our born digital documents may present. In particular, ODF brings with it a number of positive strengths:

- Multilingual support
- Can Support formula and technical notations
- Can match most of the formatting capabilities of Microsoft Office
- Underlying format structure is based on XML, and is “human readable.”
- Presently the only finalized standard for editable office documents that is currently shipping and has been vetted by an independent recognized standards body.

Drawbacks:

- Competition from Microsoft: there is concern that Microsoft, with its proposed standardization towards its own competing XML-based open document format, will garner support and adoption away from ODF and towards its own standard. This could potential call ODF’s future relevance and common usage into question.

Current Software Supporting ODF:

- OpenOffice (OpenOffice.org – free software)
- StarOffice (Sun Microsystems)
- MS-Office (via a third party plug-in under development, support not native)
    - *Note: Microsoft's official stance is they will not support ODF in their products*
- IBM Workplace
- OpenDocument PHP and Perl libraries

OpenXML: A Microsoft-supported open document standard

OpenXML - formally known as **Microsoft Office Open XML -** is a primary file format to be used by the upcoming release of Microsoft Office 2007. Microsoft has stated it will be an open standard, and has announced plans to submit it for ECMA standardization process and later to ISO. ECMA announced Dec 9, 2005 that it had accepted Microsoft's proposal to document the standard. The ECMA technical committee producing the standard is comprised of representatives from Apple, the British Library, Canon, Intel, Microsoft, NextPage, Novell, Pioneer, Statoil ASA and Toshiba.

Microsoft also published a "covenant not to sue" covering intellectual property rights it has in the new format. This covenant is very similar in wording to the one Sun provided for ODF.

OpenXML is largely a format developed in response to the criticisms originally addressed by ODF: that previous MS Office document formats were closed, undocumented and proprietary. While not compatible with ODF, it will share many of the same primary characteristics.

Capabilities:

- Multilingual support
- Can Support formula and technical notations
- Underlying format structure is based on XML, and is "human readable."
- Will be native to MS Office beginning with the 2007 release. Microsoft has stated that a design goal for its formats was 100% compatibility with the existing base of documents and formatting used by its customers.
- Will utilize MathML and Dublin Core for metadata

Drawbacks

- **This standard is not yet available to end users.** There is, at present no software which has yet shipped that supports OpenXML.
    - Consequently, our use of objects utilizing this datastream will have to be delayed until such software does, in fact, ship.
- Being a Microsoft and largely commercially-support endeavor, OpenXML has received a great deal of criticism. By its very nature, OpenXML is not vendor-neutral, while ODF does carry that distinction. Therefore, the question of whether this standard is truly "open" is debatable.

Current software supporting OpenXML:

- **NONE** as yet
- Support for OpenXML is expected to begin with the release of MS Office 2007

Adobe PDF 1.4 / PDF/A: An established standard to augment object datastreams

The first draft of an international standard that defines the use of PDF for archiving and preserving born-digital documents was submitted in 2003 to the International Organization for Standardization (ISO) for review. The choice was a logical one; although a largely proprietary file format developed by Adobe Systems, Inc., Portable Document Format files have proven to be an effective and common format for document sharing. Despite not originating as an open standard, PDF has managed to be among the first file formats to effectively transcend platforms and become ubiquitous to most end users.

The purpose then, for PDF/A, is to remove the last stumbling block to open acceptance of the format as an archival standard: its innate proprietary nature. In its current draft, the proposed PDF/A format is an open, published file standard that is based on a constrained subset of Adobe PDF version 1.4 - equivalent to format that shipped with Adobe Acrobat 5.0. This affords the digital archivist a number of advantages when selecting PDF/A in creating archival data streams.

Capabilities:

1. Full compatibility with existing PDF file formats
2. Can be opened by existing cross-platform software, most of which is already likely to be installed on end-user workstations
3. Existing PDF creation software integrates well with most software packages. A user viewing a PDF derivative file does not need to own the software that created the original document.
4. Near effortless reproduction of graphical data, mathematical equations and scientific expressions
5. Non-standard fonts and renderings can be preserved through embedding of the required elements into the PDF file.

Widespread user base and recognition helps ensure longetivity

In theory, a finalized PDF/A standard will be widely recognized and supported by agencies tasked with digital preservation. Initial drafts have been recognized by the National Digital Infrastructure and Preservation Program (NDIPP), a collaborative initiative of the Library of Congress. Further, a standards working group and a PDF/A Center of Competence has been jointly established by the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES) and the Association for Information and Image Management, International (AIIM International).

Lack of standard finalization bars recommendation of adoption

At present there has been identified one, and only one, major drawback which bars a recommendation that we adopt this standard: PDF/A has not yet been finalized. As mentioned previously, the proposed PDF/A format is a constrained subset of Adobe PDF version 1.4 - equivalent to format that shipped with Adobe Acrobat 5.0. However, there is no guarantee that the finalized standard will continue to resemble this revision. In fact, there is internal pressure within the standards committee and Center of Competence for PDF/A to consider adopting a subset to PDF revision 1.6 (the equivalent of which shipped with Adobe Acrobat 6 and 7).

At this time, the conclusion is that while PDF/A is a promising standard, we cannot presently rely on it as an authoritative archival medium because the standard has yet to be finalized. Until the draft has been ratified as-is or modified and formalized, the uncertainty surrounding the standard introduces significant risk, and cannot be the sole format on which we entrust born-digital objects for preservation.

Conclusions and Interim Solution: Rich Text Format/LaTeX with a PDF 1.4 derivative.

While we wait for a formal PDF/A standard to be realized, it may be worthwhile to include a PDF 1.4-complaint data stream in addition to at east one other agreed-upon standard as an archival master. At present, the best candidate would be Rich Text Format (RTF). Although RTF is not proprietary, having been developed by Microsoft, the format does bring with a number of compelling advantages:

1. RTF is intended to be "future proof" in that all formatting and content is human-readable using a basic ASCII-derived markup language, and does not necessarily require a specific software package to view.
2. Nearly all current word processing packages natively support the file format.
3. The standard has remained backward and forward-compatible throughout its versioning history since its introduction 19 years ago, in 1987.

Disk space requirements for born digital PDFs derived from postscript should be minimal, and would serve to augment the RTF archive, until such time as the PDF/A draft is finalized. We can then use existing pipeline methods and code to conform to PDF/A when we finally do have a standard to work with, should we determine that the final standard is suitable to commit to.

The major concern with the use of RTF, and a legitimate one, is that the format does not handle mathematical equations and scientific expressions very well. For documents containing such content, it may be necessary to include yet a third archival data stream to accommodate the needs of such content. The LaTeX2e standard appears to be a suitable, well-developed, and well-supported open source standard to adopt for such cases.

Sources:

1. Adobe Systems, PDF Reference, 3rd Edition (PDF 1.4 Specification)
2. Microsoft: Rich text Format, revision 1.8 Specification
3. PDF Tools AG – PDF/A Center of Competence
4. NDIPP: "Sustainability of Digital Formats: Planning for Library of Congress Collections" PDF/A: PDF for Long-Term Preservation.