Document Distribution Note
Release to Members: Immediate
Release to Public: 13/09/2013

DigitalPreservationCoalition

DPC – PDF/A3 briefing

University of Leeds, 13th March 2013

Introduction

The DPC hosted an invitational briefing in March 2013 to explore the potential of the new PDF /A3 standard for preservation.

These notes are intended to provide an informal briefing for members of the DPC not able to attend in person. They are a partial account published immediately after the event, and have not been validated by participants or proofed for consistency. For an authoritative and comprehensive report, readers are encouraged to contact the organisers or speakers directly or to review the slides from the workshop which will be published on the DPC website.

About the event

The portable document format (PDF) is ubiquitous, easily-produced and is widely used in a diverse range of environments. A variant of the standard – PDF/A in which 'A' stands for archive – was published in 2005. This version of the standard, also published as ISO 19005, minimises the dependencies between the contents of a file and the system on which it is rendered. This self-contained characteristic of PDF/A makes it particularly attractive for those interested in preservation. In 2008 a DPC Technology Watch Report evaluated PDF/A, commending it as a useful tool for the preservation of documents. The same report also recommended that planners should stay alert to new developments around the standard.

The latest iteration of the PDF/A standard – version 3 – was published in October 2012. This new version allows the embedding of arbitrary files which means that PDF/A3 files can be used as a wrapper and file system for digital objects. This extends the use case for PDF/A – meaning that it now has the potential to become a way of arranging, describing and encapsulating archives. However it also creates the conditions for new types of dependencies, threatening the self-contained character of the original specification.

At this DPC briefing, leaders in the PDF Association presented the PDF/A3 standard with a period of question and answer so that DPC members can better understand how it could be used in their work. An extended discussion followed in which the potential of the standard was evaluated by leading practitioners.

The workshop was designed so that participants would:

- Be updated on the latest developments of the PDF/A standard
- Have an opportunity to discuss PDF/A3 with developers behind the standard and tools that support it
- Be invited to discuss the implications of the PDF/A3 for their own preservation plans
- Be encouraged to contribute to on-going development of PDF/A and related tools

- Meet others using or considering the PDF/A family of standards within their own preservation architectures

Presentations and discussion

Gary Hodkinson (PDF Association) – Introduction to the PDF A3 standard

The PDF file is a format created by Adobe. It's a wrapper but it's not just electronic paper. Different objects can be wrapped inside it. The PDF/A standard addresses a longer term set of issues. The adobe reader is not a validator so we need a better way to validate the content. PDF/A is device or software or version independent i.e. the content is displayed consistently on any device. That's more of a challenge than you might think. PDF/A is also self-contained insofar as it contains all the components necessary to display it. The standard has to evolve. PDF/A is not 'read only' – that's a common misunderstanding. There are 3 generations ISO 19005-1, 2 and 3. The recent one was published in 2012 and it allows embedded files. PDF/A is a good option for long term archiving, especially for electronic documents. It's particularly good for archiving PDFs and it's a high quality form of the PDF. It's ideal for anything that is 'printable by you'. PDF/A can't contain active content (javascript) or encryption. No component of a PDF/A may jeopardise the consistent format of the document. Hyperlinks to external content are allowed so long as it is recognised that external content is not part of the preserved content. Fundamentally the presentation of the document must not rely on external dependencies.

There are alternatives for alternative purposes, such as TIFF for digitisation. But these have other problems. EG they cannot include embedded standardised metadata and there no unified file format is possible. Some relevant alternatives struggle to deal with the irreversible changes that are introduced by proprietary vendors. Options for stability are very strong with PDF/A.

Immediate and contentious part of PDF/A3 is that it sounds like a contradiction. Because we can attach files within the PDF/A3 there is a deviation from the 'self-containedness' of the initial versions. The key is to be clear about the archive and non-archive material that may be contained. More specifically, avoid saving archival material as an attachment. Attachments come in three types: data, source or alternative. For example you can attach a 'data.xls' as the underlying data for a diagram; you could include 'metadata.csv' to describe the document; you could include 'text.doc' as the source for the PDF/A; you could attach audio.m4a as an alternative to the text of the PDF. These three – alternative, source and data are canonical. It's possible to extend these three types of relationship though they need to be approved through the standards body. The viewer doesn't need to render the attachments. In principle, it allows any kind of file to be attached but practice there are limitations to what can be attached - .exe files for example which are problematic and which cannot be read by the Adobe reader. As far as scanned content is concerned, metadata is often the most important element so PDF/A3 allows you to store metadata in any schema you use routinely. Digitised documents are therefore always in context. For a born-digital document which is still in its active life-cycle means that you can have a single document for preservation and use at the same time. Edit the living document (say a word document) and the PDF/A will be updated as you proceed. So documents are born with obsolescence controls at the start. A question arises about the application layer needed to manage the co-ordination of the attachment and the PDF/A3. Possible use case for redaction and/or sensitivity review?

Marc Fresko (Inforesight) – Discussion paper

Much of the discussion of PDF/A3 seems to be predicated on the idea that people might use it wrongly. The fact that a standard is capable of being misunderstood or be incorrectly deployed is not an argument against the standard. Majority of preservation discussion happen in the cultural memory sector, but the majority of the problems are in other context: nuclear decommissioning or pharmaceutical for example. A couple of use cases which are familiar to record managers in these sectors show the potential of PDF/A3 and included elements that no existing technology does (elegantly) yet. Think about an email with an attachment which itself contains some embedded files. This contains multiple files with multiple dependencies and is a common use case. PDF/A3's capacity to handle content within the wrapper and although it's a complicated structure it's an elegant solution to a widespread problem.

Johan van der Knijff (KB) – Discussion paper

PDF/A3 can be compared side by side with JPEG2000. There's too little support for PDF/A3 from the OSS community … see slides for more too much else going on.

Discussion session

A wide-ranging discussion followed. Some points:

- Concern that users will think 'I have put all my attachments into a PDF/A3'file therefore I have discharged archival functions for all my content.' Not true, and not what is intended by the standard.
- Education needed about what the PDF/A3 standard means – specifically the need to specify the archival versus non-archival components
- JP-lyser tool recognised as model of the sorts of software needed. Consider PDFA3-alyser?
- PDF/A competence centre beginning work on PDF/A validator software. Needs input from DPC community. Could consider hack-a-thon to test tools or other input to committee. Need to establish broader dialogue between PDF standards developers and DPC community.
- Need for clarifying publication about PDF/A3, such as revision of TWR or possible ISO information paper
- Sense that digital preservation conversation happens in cultural memory community but majority of digital preservation problems exist in industry / regulatory sectors. Underlines need to broaden sectors represented in DPC.
- JP-lyser tool recognised as model of the sorts of software needed. Consider PDFA3-alyser?
- PDF/A3 has potential to take obsolescence out of the equation entirely. Could be the first proper case study of a preservation-ready object if issues can be addressed.
- Need to embed PDF/A3 properly within a wider risk management framework. Gradual decay of non-archival embedded objects within PDF/A3 wrapper is not necessarily a problem provided risk is properly understood.

- Most archives suffer from a gap between the formats they can actually handle and the formats that they receive. Calling PDF/A3 an 'archival standard' will make that job harder in the short term as they need to communicate a more complex message.
- Issues of scale – QA by hand is not sustainable for most archives and expensive.

About this document

Version 1	Written at workshop	13/03/2013	WK
Version 2	Distributed	13/03/2013	DPC members