TITLE: Archiving in the Networked World: PDF in 2109

AUTHOR
Michael Seadle is editor of *Library Hi Tech*. He is also a professor at Humboldt University in Berlin, Germany, and director of the Berlin School of Library and Information Science. He teaches long term digital archiving.

Archiving in the networked world: PDF in 2109?

Abstract

Paper type: research paper.

Purpose: To consider whether PDF formats are appropriate for long term digital archiving.

Design/methodology/approach: To examine how well PDF's capabilities fit eReader devices such as future scholars may be use in addition to or instead of paper printouts.

Findings: Fixity is the advantage that PDF offers for archiving, while its alternatives generally offer greater flexibility for eReader devices. The question for long term digital archiving is whether fixity or flexibility best suits the interests of future readers?

Originality/value: PDF is widely accepted as a digital archiving format and PDF documents are found in virtually every repository. There has, however, been little discussion of whether the fixed format is not in fact a long-term disadvantage.

Introduction

In 2008 Adobe's Portable Document Format (PDF) 1.7 became an international standard "ISO 32000-1*, Document management – Portable document format – Part 1: PDF 1.7*":.

> "ISO 32000-1:2008 specifies a digital form for representing electronic documents to enable users to exchange and view electronic documents independent of the environment in which they were created or the environment in which they are viewed or printed. It is intended for the developer of software that creates PDF files (conforming writers), software that reads existing PDF files and interprets their contents for display and interaction (conforming readers) and PDF products that read and/or write

> PDF files for a variety of other purposes (conforming products)." -- ISO, 2009

PDF/A, which embeds all fonts that were used in a document, had become an international standard (ISO 19005-1) three years earlier:

> "ISO 19005-1:2005 specifies how to use the Portable Document Format (PDF) 1.4 for long-term preservation of electronic documents. It is applicable to documents containing combinations of character, raster and vector data." -- ISO, 2009a

These steps have made PDF and PDF/A broadly acceptable for long term digital archiving. The advantages of PDF are manifold. Essentially all word processing programs from LaTex to Google Docs, and from Open Office to Microsoft Word can create reliable, interchangeable PDF files. PDF also accurately reproduces the formatting of documents. Open Office will read Microsoft Word files, but it can also lose nuances of the formatting, such as the type of bullets. PDF stabilizes a document's formatting across systems and across time. A PDF from 2001 looks the same in 2009 as it did in 2001 and presumably will look the same in 2109. For archiving this can be highly desirable. PDF also does a reliable job of reproducing mathematical formulas embedded in text. For any scholarly field where more than very simple mathematics matter, this is important.

Nonetheless PDF has some disadvantages that may matter in the future.

PDF and Paper

Context matters: both the social context of the users and the context of the document as an artifact. This is elementary anthropology. Contemporary users come to documents with culturally programmed assumptions about how they will read, print, and otherwise interact with a digital document. In terms of long term digital archiving an anthropologist might reasonably expect these assumptions to be different in 100 or 200 years. While no one can reliably predict how such interactions will change in the future, contemporary experience with digital documents offers some important clues.

The document server of Humboldt-Universität zu Berlin recorded over 94,000 accesses to PDF documents in August, 2009. (HU, 2009) PDF represents the majority of the full text documents on the document server, which is the largest in Germany and 35th largest in the world according to the "Ranking of Web Repositories" (Cybermetrics, 2009). The overwhelming majority of the PDF documents appear to be in A4 format. In other words, this corpus of texts looks as if it were on standard German (international) A4 printer paper, which is only slightly longer and narrower than US letter-size (8.5 by 11 inch) paper.

The same emphasis on printer-paper format holds true for PDF documents on the Social Science Research Network (SSRN), which is ranked as number three in size but number one overall based on factors such as full-text availability. Part of the reason for the use of A4 or letter-size paper images in PDF is that many publishers allow only pre-prints to be put into repositories. SSRN has a component explicitly for preprints in "an Electronic Paper Collection currently containing over 205,000 downloadable full text documents in Adobe Acrobat pdf format." (SSRN, 2009)

While these statistics represent only a modest sampling of what goes into repositories worldwide today, they show an important trend. Since authors themselves generally submit documents to repositories, the dominance of PDF in printer-paper formats suggests that these authors regard those sizes as normal. It is what they and other people are accustomed to with both printers and photocopy machines. Those accustomed to photocopying of paper journals in libraries also end up with A4 or letter-size paper, regardless of the physical size of the journal page itself. It is simply the norm today.

The paper sizes A4 and letter-size are comparatively recent. Both date from an urge for standardized production in the early part of the twentieth century. Today A4 and its relatives dominate as ISO 216. Paper that is letter-size remains idiosyncratically north American. (Kuhn, 2006) Before this early twentieth century standardization, sizes varied considerably, as anyone knows who has looked at correspondence from earlier eras. Folders, binders, filing cabinets and other storage mechanisms exist specifically for these paper sizes, which makes changing them economically costly and therefore less likely, even over long periods. This argues well for the use of PDF for archiving -- assuming paper remains a significant medium for reading in 100 or 200 years.

Formats for eReading

Three years ago most librarians I asked insisted that paper would be the standard reading medium indefinitely. In the last year or so that certainty has wavered somewhat. The reason has to do both with the onslaught of publicity around eReader devices such as the Kindle, the Sony, and the iRex, and that fact that the sales of digital content for these devices has done well. I have used both the Sony and iRex. It is relatively easy to put PDF documents on them, and it is now possible to put PDF on the Kindle 2 .0 for free via a number of fairly simple work-arounds [1]. I have not tested the results on the Kindle. I have on the other two.

The Sony PRS-505 has a relatively small screen (newer models have larger ones). In many respects this is an advantage, since it means that the Sony is as small as a thin paperback book and fits easily into a jacket pocket. I can read a PDF on the Sony at a resolution which retains all of the formatting, as long as I use the built-in option to rotate the image 90 degrees and, in effect, read the work sideways. An A4 page takes two screens to display. For an extremely fast reader this could be annoying, since page turns require a small pause, perhaps a second or a fraction more. I find reading PDFs on the Sony quite acceptable, as long as there is good light on the page (all of these devices use reflected light just like an ordinary printed page). The iRex device is larger, more the size of a hardcover book, and can display a whole PDF page, though in a smaller font that would be on A4 paper. The iRex also allows making notes on the text. The Sony only allows bookmarking pages. Since the Kindle 2.0 has a larger format, reading a PDF is probably similar to an iRex.

While a PDF in A4 format will work on contemporary eReader devices, it is not always a comfortable experience, especially when authors choose smaller typefonts. In low light situations I must use the magnification options to read a PDF on the Sony. In the magnified modes the Sony uses the text behind the PDF image and shows broken lines. The text is quite readable, but the original formatting is lost. Tables especially become a problem. I have not tested whether magnification has the same consequences on the other devices. Although the loss of formatting under magnification is not ideal, it seems far better than carrying a thousand pages of masters theses in a backpack in order to grade them while at a conference or on a trip. There is good reason to think that the software and hardware will improve.

eReader devices are still comparatively rare. Most online reading of PDFs today is done on laptop or desktop computers whose screens are, perhaps for both historical and ergonomic reasons, ill adapted to either A4 or letter-size paper. Reading a PDF on these devices is easy as long as there is a good scrolling mechanism, such as on the MacBookPro, where two fingers on the touchpad gives instant control. Those using the page-down / page-up buttons have a less comfortable experience, since the page jumps in ways that make it easy for the eye to lose its place. These are minor awkwardnesses, however. In general it is reasonable to claim that even if people give up printing paper versions of PDFs and read them instead on digital devices, they could do so easily now, and probably more easily in the future.

eReading Habits

Thus far my argument has been that PDF in standard paper formats is probably OK for eReader devices, but is it ideal? If, in the future, eReader devices do replace paper, then limiting pages to an arbitrary length set in the twentieth century may seem like an unnecessary artifact of a bygone physical technology. Hard statistics about online reading are difficult to find and at least one article in this issue offers evidence for an ongoing preference for paper -- in some countries. (van der Velde, 2009) Nonetheless even scholars who say they print everything often admit to using within-document search functions and to reading enough of a work online to decide whether to read the rest in print. Specific objections to on-screen reading often have to do with eyestrain from the backlighting on conventional computer screens or with the awkwardness of holding even a moderate sized laptop compared to a sheaf of papers. The first generation electronic-paper-based eReader devices have eliminated both of these major objections. No one can say with certainty when or whether paper will vanish from use as a medium for reading, but the amount of online reading seems likely to increase significantly in the next century or two, which makes some thought about the consequences worthwhile.

If paper use does decline as eReaders improve, will eReading devices adapt themselves to paper-sized displays? There is certainly evidence that early printed books imitated the format and layout of manuscripts. To some extent the iRex and perhaps the Kindle formats have done this, but that could also be transitional. There is no standardization imperative for a single page size in the electronic reading world. Even paper books come in a wide range to sizes, a fact well known to librarians trying

to fit them all on a single shelf in call-number order. Our inability to predict the future should extend also to assumptions about paper and paper-based formats remaining as a reading medium. Flexibility is critical for long-term planing and should be inherent in archiving assumptions and standards.

PDF Alternatives

What are the alternatives to PDF? While no product retains the exact formatting as well as PDF, many retain key elements. The software on the Sony eBook reader will format a work in Rich Text Format (RTF) so that it looks good and fits the screen size perfectly. RTF has a history that dates from the late 1980s and word processing packages like Microsoft Word and Open Office continue offer it as an alternative. LaTex, which is widely used in computing, physics and other engineering and natural science areas has no automatic RTF option, though creating one should doable. RTF can handle embedded images, but (in my experience) tends to do it badly. The various eBook formats such as MobiPocket or EPUB are also thinkable alternatives. XML using a TEI (Text Encoding Initiative) document type definition would be ideal, if a good reader existed. The eReader market is still far too new and too open to guess which has the best chance of prevailing. Copyright considerations may well make a format that works with Digital Rights Management protection software a factor.

Depending on how much formatting needs to be retained, plain ASCII could be an alternative. ASCII would be unacceptable to anyone who wanted to see a book or article in some semblance of its original form, but such people tend to be a minority. Most readers care primarily about the intellectual content of a work. Evidence for this comes from the way publishers reissue ninetheen century classics. Authors like Dickens first serialized their novels in newspapers or magazines, then published them in multi-volume editions. Today the novels appear in bookstores in single volumes with footnotes to explain obscure terms, modern typefonts and contemporary orthography. The chapter structure generally remains, but the exact paragraph structure may not. The number of words and letters appear on a line will almost certainly vary with typefont changes.

The ASCII texts in project Gutenberg [2] offer a good example of what happens when formatting is reduced to its ASCII basics. For the book historian something certainly is lost, but something is also gained, because it is easy to take the Gutenberg texts and

transform them into formats for eReading devices. Gutenberg has offered HTML and Plucker formats for some time. Both have allowed standard text formatting including links, tables and embedded illustrations. Gutenberg has also started experimenting with other formats, including EPUB, MobiPocket, and QiOO Mobile. Gutenberg's formatting appears largely to derive from the original ASCII version with added links to chapters, paragraph markers, and sometimes italics, bolding, or em-dashes. The formatting makes no pretense of reproducing the original.

All of these formats allow more compact files than PDF, but that no longer matters on contemporary storage devices. Their real advantage is that they automatically take the shape of the available screen space on eReading devices. In other words, flexibility is their advantage in contrast to PDF formats, whose value comes from its fixity. One is not necessarily better than the other. They serve different purposes. The question for long term digital archiving is: which purpose best suits the interests of future readers?

Conclusion

What is the value of preserving the exact format of the thousands of pre-print articles or unpublished theses and dissertations in contemporary repositories? For some format elements, it is high: for example, an exact and unchanging rendering of mathematical expressions, something that ASCII, HTML, and similar markup-languages tend to do badly and irregularly. For non-mathematical works, especially those consisting almost entirely of simple text and tables, the exact typefont, font size, width of margins and other page layout features arguably have no value to the vast majority of future readers. They may merely use PDFs to strip out the text and reformat it in ways that fit their reading media, which might even be paper of a different size and shape.

Preserving digital works in exactly today's shape and form appears rational to people who grew up in a world of fixed paper media. It may not make sense in the future.

Notes

[1] See: http://ireaderreview.com/2009/02/09/kindle-2-pdf-faq-pdf-kindle-20-conversion-questions/
[2] see: http://www.gutenberg.org/wiki/Main_Page

References

Cybermetrics Lab (July, 2009), "Ranking of World Repositories". Available: http://repositories.webometrics.info/top400_rep.asp

ISO [International Organization for Standardization] (2009) ISO 32000-1:2008 - Document management -- Portable document format -- Part 1: PDF 1.7 [Internet]. Available from: <http://www.iso.org/iso/catalogue_detail.htm?csnumber=51502> [Accessed 17 September 2009].

ISO International Organization for Standardization (2009a) ISO 19005-1:2005 - Document management -- Electronic document file format for long-term preservation -- Part 1: Use of PDF 1.4 (PDF/A-1) [Internet]. Available from: <http://www.iso.org/iso/catalogue_detail?csnumber=38920> [Accessed 17 September 2009].

Kuhn, M. (2006) A4 paper format / International standard paper sizes [Internet]. Available from: <http://www.cl.cam.ac.uk/~mgk25/iso-paper.html> [Accessed 14 September 2009].

Social Science Research Network (SSRN) (2009), Leading Social Science Research Delivered Daily. Available: http://www.ssrn.com/

van der Velde, W. and Ernst, O. (2009), The Future of eBooks? Will Print disappear? An End-User Perspective, *Library Hi Tech*, Vol. 27, No. 4.