



Information Overload

Issue 19.1, May 2004

Welcome:

The subject of electronic archiving continues to dominate the discussion boards and list servs, and will continue to do so until such time as a definitive answer is found. We have taken the decision to issue a follow up to the electronic archiving discussion paper that was issued in March of this year. I would personally like to say a very big thank you to everyone who took the time to comment on the newsletter, whilst we haven't used every comment we have chosen those which we feel expand on the issues raised in the original piece – please note - names have been withheld to protect the innocent.

You may wonder why we have issued the newsletter out of sequence and out of synch with the normal publishing schedule – well to be honest we have some interesting issues coming up and we didn't want to delay any of the topics any more than we have to.

Comments as always can be sent to training@iea.com.au.

We hope you enjoy reading.

Lorraine Bradshaw
Marketing & Training Coordinator
Information Enterprises Australia P/L

Electronic Archiving: A Follow Up

PDF: Adobe's Portable Document Format appears to be the most widely accepted and used format (after Microsoft) for the transmission and "archiving" of electronic documents. However, the merits of whether this is the best format to use continue to be debated, and will continue to be debated until a definitive answer is found to the very real problem of what do we do with electronic documents that we want or need to keep, but do not need to refer to on a regular basis.

In our white paper on electronic archiving, we discussed the issues surrounding the "archiving" of documents into the PDF format. And as we tried to show, using PDF as an archiving medium was flawed on more than one count.

For example:

- A PDF is a snapshot of how a document "looked" at a particular point in time. Like its other electronic counterparts, a PDF document may also be overwritten or deleted (accidentally or otherwise). Even with security measures in place, the person who attaches the security to the record can remove it. Throwing into doubt the records reliability, authenticity and accuracy.
- A PDF does not contain the 'metadata' of the original document. For instance, in a "Word" document you can track changes, and know which machine was used to make them. If an organisation insists on strict password control over individual pc's, then the changes may also be linked back to an individual.
- Where are the documents stored once they are converted to PDF? Will they be kept "live" or will they be "archived" onto a different part of the server, or an entirely new server. Or will they be archived onto a different media entirely, for instance CD-R or DVD-R. What happens when the archive server becomes

obsolete and needs replacing? Who will check to make sure that all the existing PDF documents can be read with the new software and hardware?

It appears that we are not alone in our concern on this issue:

JH of Australia said "I have been arguing precisely the same thing for the last couple of years but, at least in the exalted halls of academe, PDF's are seen as the answer to a multitude of prayers, primarily because of the security features. Still, I'll continue to archive everything of importance the way I have been (generally, in two different formats) and some years down the track I am sure I'll have the last laugh or, at least, the last readable documents!"

And in a conversation with HQ also of Australia, "We are looking to raise awareness of the issues surrounding digital archiving, and the management of records throughout their electronic "life". We use PDF at the moment, but we are also preserving the original bit streams just in case. Having just returned from a Conference on Digital Preservation in the United States, there were a lot of discussions regarding the use of PDF and JPEG 2000 as container based standards – these can contain embedded XMP metadata, JPEG 2000 can also store graphics and videos. There was also discussion over the continued use of microfiche, paper based records being sent to "Iron Mountain" and "Locked Servers" across the world. These locked servers would be mirror sites held in multiple locations to ensure survival of records in the event of major disasters. HQ also said that whilst they were using PDF at the moment – it was because it seemed to be the best option available at the current time. "

Thankfully it appears that work on an archival version of PDF is on the way. JT of the UK asked *"There was one question I wanted to put back to you; are you familiar with the activity around PDF/A? The last couple of times I've spoken with our friend Rich Lysakowski, he indicated that there was commitment within US Government departments to adopt this as an archiving standard. It is being developed as an ISO standard."*

The reason why the United States Government is pushing the electronic community into creating an international standard of the popular format is a simple one. Two of the largest bankruptcy filings in U.S. history – Enron Corp and Global Crossing, produced a record number of PDF documents, which federal governments have to archive and preserve.

Organisations such as Eastman Kodak Co., Global Graphics Software Ltd., IBM Corp., PDF Sages Inc, Xerox Corp and of course Adobe are looking to create a new international standard that will help to solve some or all of the problems of archiving electronic documents in PDF format. It is hoped that the new version will be available by early 2005. It is also interesting to note that Adobe has relinquished all proprietary rights in perpetuity to version 1.4, the specification on which PDF-A is based. Whilst this does not make it "open source software" it is certainly a step in the right direction.

Whilst XML (Extensible Markup Language) is another strong contender for archiving electronic data, it is recognised that PDF-A is winning the battle simply because it retains the "appearance" of the original document – an important point when considering litigation and case law. "For records of legal proceedings, the position of paragraphs and footnotes by reference to the page number on which they appear in a printed document is crucial for understanding because attorneys rely on positional reference when they present their arguments." (1)

Another step in the right direction is the move to introduce tighter security measures for documents created as a PDF. John Landwehr, Adobe's group manager for security solutions and strategy was not willing to divulge details relating to particular cases of document tampering – but he did say that "document spoofing represent a growing problem for government and corporate offices." (2)

Just as the Sarbanes-Oxley Act was introduced by the US administration to prevent cross selling of services by auditing firms to their clients, and thus prevent further conflicts of interest (as in the case of Arthur Andersen and Enron Corp), so document spoofing has made organisations sit up and take notice that not all is well in PDF archiving terms. A case of "oops the horse has bolted, we'd better shut the door." My question is - How can we ensure a document's integrity in the face of litigation? Complying with the Sarbanes-Oxley Act for instance requires that an organisation should produce, on request, authentic and reliable records and all supporting documentation. And states:

"Anyone whoever corruptly – (i) alters, destroys, mutilates, or conceals a record, document, or other object, or attempts to do so, with the intent to impair the object's integrity or availability for use in an official proceeding; or

(ii) otherwise obstructs, influences, or impedes any official proceeding, or attempts to do so, shall be fined under this title or imprisoned not more than 20 years, or both." (3)

Whilst Adobe are attempting to protect the documents integrity with the use of a "policy server" (ibid: 2) more emphasis needs to be placed on the use of digital signaturing technology and Public Key Infrastructures (PKI's) to protect electronic documents. As we discussed in our original paper, electronic records are easily altered, edited or deleted and therefore cannot be relied upon, should we need to produce electronic records in a court of law.

To add to our woes as records managers, librarians and archivists there is yet another twist in this story. Earlier this month, the Sunday Times advertised "The Next Big Thing." Called the Blu-ray disk – it can hold between 23Gb and 54Gb of data (approximately 12 hours of standard TV recording), and a massive increase on the standard CD or DVD technology (700mb and 4.7Gb respectively). Of course in order to utilise this new technology you will need to purchase a new "recorder" which is being produced by Sony. With the price of DVD and CD burners dropping through the floor there is little surprise that a replacement wasn't around the corner. However, what makes this technology incredible is the fact that the disks are made of 51% paper fibre. Toppan printing (joint producers of the new technology) says that "the fact that you can easily cut it up with scissors offers foolproof data security when it's time to bin that old disc." (4)

Is anyone else concerned about the ramifications this technology has on the record keeping industry worldwide? Because apart from the obvious concerns regarding fraud, we now have to go back to worrying about those cellulose munching pests eating their way through our digital archive as well as our paper based records.

And finally RD of the USA had this to say:

"Lorraine, talk of synchronicity-I had been asked to help an editor with an analysis of new products in the chem informatics area.

The reply went as follows. We both ended on the same note, or quote:

"As Keats said (I think); 'The best lack all conviction, while the worst are full of passionate intensity' The use of such worn quotations and other overused clichés sums up the situation.

The three most important things in the storage/retrieval of analytical data are "standards, standards, standards".

Many of the products are novel, pointing towards a future. But "the future is a foreign country, they do things differently there" (Arthur Clarke ???).

What users need, and would prefer, is a lingua franca format that eschews the need to understand the presence of complex translation programs, plug-ins, and lost data fields.

Presently most vendors profess to a standard storage format, but it is not universal. Yet digital photography enthusiasts enjoy the ability to paint their photos from a few standards, and mix simple programs.

For many years it has been evident that a gap exists between what users want, and what vendors are willing to provide. The community does not intra-communicate.

Plus sa change, plus c'est la même chose."

Thank you to everyone who took the time to reply. Your comments are very much appreciated.

Notes:

(1) Olsen, Florence; Archivists praise PDF/A; Federal Computer Week March 10, 2004.

<http://www.fcw.com/fcw/articles/2004/0308/web-pdf-03-10-04.asp>

(2) Olsen, Florence, Fortifying PDF documents; Federal Computer Week, May 10, 2004.

<http://www.fcw.com/fcw/articles/2004/0510/tec-pdf-05-10-04.asp>

(3) The Act can be viewed in its entirety at:

<http://news.findlaw.com/hdocs/docs/gwbush/sarbanesoxley072302.pdf>

(4) The next big thing; STM (The Sunday Times Magazine) May 9 2004, p4

A Thought to Ponder:

"We live in a rainbow of chaos"

Paul Cezanne

Your comments and suggestions on the subject of this newsletter are most welcome. Or if you would like to see other issues covered in future editions, please email training@iea.com.au

Please feel free to pass on this newsletter to your colleagues' friends and associates. To subscribe they should send an e-mail to training@iea.com.au with "subscribe newsletter" in the subject line.

If you would prefer not to receive this newsletter, please send an email to training@iea.com.au with "unsubscribe newsletter" in the subject line. If you have any suggestions as to what should be included in future editions, then please send an email to training@iea.com.au.