# CHALLENGES OF ARCHIVING ELECTRONIC RECORDS: THE IMMINENT DANGER OF A "DIGITAL DARK AGE"

**Richard Wato**

**Kenya National Archives and Documentation Service**
Email: KNArchives@kenyaweb.com or r_wato@hotmail.com

**Abstract**
*Modern society is faced with the glaring possibility of a future 'digital gap' where the electronic records preserved from yester years could not be retrieved due to a number of technical and policy factors. Chances are that there will not even be any paper surrogates for those electronic records as the practice has lately been the substitution of paper records with electronic ones without due consideration of the longevity of the latter. Computer users who normally look at the current use of electronic records without necessarily considering the preservation issues that need to be considered if these records are to be available for future use have accelerated this practice.*

*This paper looks at the challenges that are likely to face the archivists in their endeavour to preserve electronically generated records. It also suggests issues that the archivist may consider in the preservation of these records. Archivists around the world have been giving warnings that computer files may survive long but the equipment to make sense of them may not. This era could become a "digital dark age" - a part of its collective memories forever lost.*

## Introduction

National Archives are responsible for providing guidance and assistance to government agencies on the creation, maintenance, use, and disposition of government records. Government agencies on the other hand are responsible for ensuring that their records are created and preserved in accordance with the legislative provisions of the national archives. Records generated electronically, such as electronic mail (E-mail) messages, word processing documents, spreadsheets, databases, images and World Wide Web pages, present a preservation challenge for archival institutions and agencies because these technologies are new and changing very rapidly. Also, the sheer volume of these records is mushrooming.

Records that have immense historical, legal, social, personal and economic significance are, therefore, at the mercy of the elements, obsolete technology, the push for a paperless society and ignorance. There is, therefore an urgent need for the archivist to critically look at the available options with the aim of coming up with a long-term solution in the preservation of today's digital heritage.

\

**The Digital Dark Ages**

Historically, the 'Dark Ages' started with the fall of the Western Roman Empire in AD 476. The conquering forces destroyed many fine buildings and works of art that had existed during the Roman times. During the 'Dark Ages', knowledge survived only in monasteries. The period was so called because we have almost no information about what happened then. This was not because very little ever happened, but because no records survived to tell us what did happen.

In the 15th century technology came to the rescue of the written heritage that had survived the so-called Dark and Middle ages. That technology was the printing press that was invented by Johann Gutenberg in the early 1450s. The printing press arguably made possible the Renaissance, the Protestant Reformation, the Scientific and Industrial Revolutions, and the technology-rich society we know today. It was actually declared as one of three greatest inventions of all time (gunpowder and the magnetic compass being the other two).

The analogy of the invention of the printing press and the explosion of the Information Age as we know it today is relevant in that during the Dark Ages the conquerors of the Roman Empire did not realize the importance of the intellectual works preserved by the Romans and for that reason they went ahead destroying everything in the name of destroying a civilization. Likewise, the Information Technologists of today do not seem to get into terms with the archivists is as far as preservation of electronic records is concerned. In this regard, therefore, unless comprehensive collective actions are taken, we are faced with the problem of losing our electronic records. This would inevitably lead to a digital dark age where the electronic records we are creating today will not be accessible by future generations.

Kuny (1997), a consultant with the National Library of Canada, struck a prophetic warning when he wrote that,

> as we are moving into the electronic age of digital objects it is important to know that there are new barbarians at the gate, and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever. We are …living in the midst of digital Dark Ages; consequently, much as monks of times past, it falls to librarians and archivists to hold to the tradition which reveres history and the published heritage of our times.

This paper has no intention of painting a dark future on the preservation of electronic records. On the contrary, the aim is to put emphasis on the dangers that we are already living with and the impending future ones unless something is done urgently. Those who have been using computers for a number of years may confess to having lost some of their electronic data for one reason or another, sometimes not as a result of technological obsolescence, but mostly because they could not remember the file name or the directory where they saved that document. This problem is not limited to technology novices but also those who consider themselves experts.

## Traditional methods of preserving records

Traditionally it has been relatively much easier to preserve records when they are stored in archives and other conventional repositories. Even poor quality newspapers are known to last a few decades while high quality paper can last 500 to 600 years. Paper may seem old-fashioned, but one can be reasonably sure it will still be legible in many years to come. The same cannot be said of a CD or a floppy disk.

While paper documents are vulnerable to physical degradation, electronic documents are not only more so, but are also in jeopardy from the inevitable obsolescence of the hardware and software necessary to store and retrieve them. In the home page of the website for the National Archives and Records Service (NARS) of South Africa one will find at the background some image. This is actually a photographic replica of rock painting done thousands of years by the San people of Southern Africa (National Archives and Records Service of South Africa 2003). Such painting did not require rigorous preservation efforts to last that long. For electronic records to last even a couple of decades, proactive efforts are required to preserve them.

Looking around our offices, one will find many paper records in very poor state of preservation despite the fact that we have time-tested methods of preserving paper records. The digital technology on the other hand is something we are only starting to learn, and thanks to planned obsolescence, before we learn to use one aspect of a software system, a second, a third and even a fourth generation of the same is already with us. This makes it quite hard to cope with the changes. The people who were expected to manage paper records but never did so properly are still the same ones who are now expected to manage the electronic records. Are they going to do it better? The answer is a definite NO!

*Definition of the term 'archives'*

Archivists and information technologists have always conflicted on the use of the term 'archives'. To the information technology (IT) sector, archives simply means those computer files that are no longer needed for current use. Such files are normally removed from online storage media and put in offline media such as magnetic tapes. The idea is to decongest the online media thus relieving it for use with current files. The use of the term 'archives' in IT, therefore, only means removal or transfer of the files from one storage media to another. In this case there is no emphasis whatsoever in the preservation of the 'archived' records. Even if there was such emphasis, the archiving media may not last long enough to ensure continued access to these files.

To the archivist, on the other hand, 'archives' refers to those records with enduring value that needs to be preserved permanently for posterity. The archivist, therefore, requires such records to be in well-tested media that guarantees long-term preservation of the materials.

This conflict of definition from the outset may seem insignificant but given a closer

look one will appreciate that it is the information technologists who develop both the hardware and software needed to create, retrieve and store the information that the archivist is expected to preserve in perpetuity. For this reason, unless there is a clear and mutually acceptable definition of the term between the two parties, IT experts will always develop tools that meet their own 'standards' with total disregard of the fact that their so called standards are far below the expectations of the archivist.

Due to the foregoing reasons, IT experts must of necessity develop a fuller appreciation of the disciplines and practices of record managers, archival scientists, and library and information scientists since these professionals are the guardians of corporate accountability. They are also important gatekeepers for intellectual capital, and levers for getting the most out of it. Information technologists must design acceptable technology solutions that meet the needs of these professionals.

Archivists and other information managers should on the other hand adopt a multidisciplinary approach in their training to include much more of IT because that is the only way they could understand the working of this technology. Once the archivist has acquired enough IT skills, he will be in a position to consult with IT specialists in the design of a records management system that will truly adequately address long-term preservation issues.

**Information communication technologies and records management issues**

Most countries in Africa today have developed national information and communications technology (ICT) policies. While this is a step in the right direction, it is disheartening to note that only a few of these national ICT policies have addressed records management issues. This is dangerous because ICT will generally lead to creation of more electronic records and if the proper care of these records is not addressed, there is a possibility of such records disappearing without trace. Kenya, while late in formulating a national ICT policy, was lucky to have learnt from the mistakes of others. Two members of the Kenya National Archives and Documentation Service were actively involved in the formulation of the policy where they ensured that records management issues were addressed properly.

The ignorance of records management issues to Information technologists came out clearly during the Kenya ICT conference when one participant from the Communications Commission of Kenya (CCK) said that when their e-mail server fills up, they simply delete the old e-mails to create room for current ones. While deleting these e-mails, they are not subjected to appraisal in order to preserve the valuable ones. This mistake usually comes out of ignorance of the fact that e-mail messages used to conduct official business are official records. But this example would have set a very bad pattern particularly coming from the organisation that is supposed to formulate policies on electronic communications in Kenya.

**Rapid planned technological obsolescence**

Developers of IT hardware and software seem to be competing with themselves to introduce into the market new models and versions of their products every so often.

While this helps the companies to make more profits since the users have to pay up for the upgrades, it has not helped in the search for a stable digital archiving solution. Currently there are no stable electronic storage media that can be considered archival. The main problem surrounding the preservation of authentic electronic records is that of technology obsolescence. As changes in technology continue to increase exponentially, the problem arises of what to do with records that were created using old and now obsolete hardware and software. Unless action is taken now, there is no guarantee that the current computing environment and its records will be accessible and readable by future computing environments.

Since technology standards are evolving at an enormous rate, the type of storage media has to be kept up-to-date. While CD-ROM's might survive and store data for a hundred years, (this is what some manufacturers claim but it has never been proven) chances are that no drives capable of reading CD-ROM's will be available by that time anymore, as new forms of storage media will by long have replaced it. New media for storing digital information rapidly replace older media, and reading devices for these older media become no longer available. We just have to imagine how many types of storage media we have seen passing in the last 30 or so years, starting from punch cards, old magnetic tapes, various types of hard disks, to 8 inch and 5¼ inch floppy disks. Indeed, technological obsolescence represents a far greater threat to information in digital form than the inherent physical fragility of many digital media.

Here are some awesome examples of instances when electronic records stored in yester years could not be retrieved or were lost altogether: -

- The Pentagon irretrievably lost all but 36 of the 200 pages of its brief desert adventure during the first Gulf War. Half of the missing files were wiped out when an officer at the Gulf War Headquarters incorrectly downloaded some games into a military computer.
- NASA scientists who recently tried to access material on the 1976 Viking mission to Mars discovered that 20 per cent of it has simply vanished and the rest is going fast. The data lost from the Viking Mars mission was trapped on decaying digital magnetic tape, forcing NASA to call mission specialists out of retirement to help the agency reconstruct key data. This led Jeff Rothenberg (1998) to quip, "Digital information lasts forever, or five years - whichever comes first."
- During the United States 1960 population census, the Census Bureau retained records for its own use in what it regarded as ‘permanent’ storage. In 1976, the National Archives identified seven series of aggregated data from the 1960 Census files as having long-term historical value. A large portion of the selected records, however, resided on tapes that the Bureau could read only with a UNIVAC type II-A tape drive. By the mid-seventies, that particular tape drive was long obsolete, and the Census Bureau faced a significant engineering challenge in preserving the data from those particular tapes. By 1979, the Bureau had successfully copied onto industry-standard tapes nearly all the data judged then to have long-term value. The Committee on the Records of Government later proclaimed that, “the United States is in danger of losing its memory” (Arnita & Cantelon 1993). It is observed that when the problem came up, there were only two machines in the world capable of

reading those tapes.

- The computerised index to a million Vietnam War records was entered on a hybrid motion picture film carrier that can't be read any more;
- Australia's earliest detailed seismic survey results - worth millions to mining, mineral and exploration interests are in jeopardy as much of the data is stored on magnetic tape and there is only one machine left in the world, in London, which can read it.
- In 1964, the first electronic mail message was sent from the Massachusetts Institute of Technology, the Carnegie Institute of Technology or Cambridge University. Today the message does not survive and so there is no documentary record to determine which group sent the pioneering message. Contrast this with how much we know about the first telegram (now digitised and on the web) or telephone message.
- Satellite observations of Brazil in the 1970s that were critical for establishing a time-line of changes in the Amazon basin are also lost on the now obsolete tapes to which they were written.
- In June 2002 a Norwegian literary museum admitted losing access to their catalogue system after the database administrator died taking the password with him. The museum put out a radio call for hackers to help crack the code.
- In 2000 the University of California, Berkeley published a study showing that printed content represents only 0.003% of the world's total information most of the remainder is stored digitally (Emberton 2002). If that figure is correct, almost our entire output as a society is entrusted to one of several Microsoft operating systems and disks with some twelve-month limited warranties.

The above examples are just but the tip of the iceberg. Many cases are unreported at our individual and corporate levels.

**Some possible solutions**

Archivists have been trying different methods to preserve electronic records. One such method involves moving much of their information over to open source formats, so that the information is not lost as the formats become outdated. With an open format, even if the creators of the format are long dead, someone can read the specifications and create a new program to view the information.

Below are some of the methods that are used in the painstaking endeavour to preserve electronic records.

*Printing hard copies*

This simply means downloading the digital document and printing off a copy that is accordingly filed in the normal filing system. This method is very commonly used nowadays especially with e-mails. The method, however, negates all the flexibility and positive features of the digital version and is hopeless in the case of dynamic

multi-media formats. Such multi-media objects cannot be meaningfully printed at all, or would lose most of their uniquely digital attributes. Printing is, therefore, no answer to preserving electronic records since even doing so would inevitability create another problem of managing the paper records. The only solution is to come up with a digital solution that is tested and found to work.

*Computer 'museums' or 'reincarnation'*

In this method obsolete hardware and software systems are preserved, so that old digital documents can be opened and read using the programmes, operating systems, and equipment on which they were originally created. This approach is cost-*in*efficient, location-bound, improbable and still fails to address the issue of making digital files compatible with current operating environments. This is because the digital media containing the original software will eventually degrade beyond repair.

*Migration*

Migration is currently one of the most widely adopted short-term approaches to digital preservation, yet it is also the one that appears to attract the most criticism. Digital Migration is defined as the transfer of record(s) from one hardware/software configuration/platform to another. A simple example of this would be the migration of a record from one version of Microsoft Word to a latter version; a more complex example would be the migration of a record from a Macintosh OS to a Windows OS.

Migration is normally criticised because its results are often unpredictable due mainly to a lack of testing and documentation. When new software is brought out, it is common for many people to simply refresh their documents by recopying. This often results in the loss of information, whether record content, format, behaviour, or appearance. The new application reads the record in a different manner from that in which it was designed to be read, and during the migration process, some processing instructions, content, and functionality, may be lost or even gained. The loss of this information varies, depending on the extent and nature of the migration performed. Migration results are difficult to predict, unless a substantial amount of work is done in advance on the source and target format specifications. Migration can affect a records status as authentic, and any record which is preserved must be preserved authentically, otherwise its meaning and validity cannot be assured.

The complexity of the migration process usually depends on the nature of the digital resource, which may vary from simple text to an interactive multimedia object. Converting data to another software format entails a loss of functionality. Cross-references, indices, interaction mechanisms, automatic updates, formulas and the like will usually be lost during the conversion stage, since the particular standard file format will not support the same functionality as the original software. Furthermore, the authenticity of the original object is thereby corrupted. In other cases, as for example for some forms of interactive art on the Internet, converting to another format without losing all of its characteristics is not possible at all.

For example, when Word'97 was released, documents created and saved in a Word97 format could not be accessed via Word'95. Thus, when a Word'97 user sends a Word'95 user a document, the latter is unable to open, view, or print the Word'97 file unless one gets Word'97 converters. Converters will obviously lead to loss of features that did not exist in the older formats.

When considering migration one needs to address the issue of compatibility. Backward compatibility of software means a newer software version can read and process files created by an older software version. Forward compatibility of software on the other hand means an older software version can read and process files created by a newer software version. People most often desire backward compatibility of the software to read and process older files, since the software is continually updated. Forward compatibility of the software is less frequently needed, only by people who fall behind in software versions, e.g., trying to read an MSWord'97 file with MSWord'95.

*Emulation*

Emulation refers to the process of mimicking, in software, a piece of hardware or software so that other processes think their familiar environment is still available in its original form. Digital documents can, thus, be kept without being altered, maintaining their integrity and original look-and-feel. Together with the documents the access software is archived. Such systems are to be run on a future system that simulates a contemporary computer environment without the software noticing, giving access to the stored digital documents.

The theory behind emulation as a digital preservation approach is that digital documents are inherently software-dependent, regardless of their format. Emulation proposes bypassing the problem of hardware/software obsolescence by enabling the recreation of the old software and the environment needed to run it inside of new and future hardware. By preserving not only the record but also the software on which it was written and originally intended to run, the record will not undergo any changes and its preservation and authenticity can be assured.

Theoretically emulation is the most stable model and a conceptually clean solution. In fact, if preserving the original functionality and recreating the look-and-feel of a document is a prime objective, it is the only reliable way. Emulation has attracted similar criticisms to migration, on the grounds that it can be costly, highly technical, and labour-intensive. The criticism is not always justified as there is currently no tried and tested specific methodology for emulation, the future costs cannot yet be predicted and may or may not turn out to be less than for migration.

*Format Standardisation and "System Neutral Formats"*

Standards are key to successful migration strategies. Since the 1960s more than 200 data-storage formats have come and (mostly) gone, while magnetic tapes and hard disks for the storage of digital information have been offered in a variety of sizes and densities. Proponents of standards as a solution to preservation naively imagine that standards will last forever. Conventional wisdom, however, shows that

no computer technical standards have yet shown any likelihood of lasting forever, indeed most have become completely obsolete within a couple of software generations.

Most systems developers do not even support standardization, citing the following reasons:

- Cost - it takes scarce time and resources away from competitive features to define and implement standards. They argue that standards are always a cost-center that do not bring in any profit,
- Poor current acceptance of most standards as no one wants to be first to say goodbye to their proprietary technology,
- Short lifetime of most standards especially if no one really commits to them,
- Standards will always compromise some areas that the vendors dearly believe in,
- The not-invented-here syndrome where everybody wants to be the standard while nobody wants to follow someone else's standard.

Notwithstanding the foregoing, some developers have come up with standards that have shown significant possibilities of success: -

*Portable Document Format (PDF) archive*

The storage format that comes closest to a universal standard for high-quality electronic records is Adobe System's Acrobat Portable Document Format (PDF) developed by Adobe Systems. PDF is a "container format" that permits capture, publishing, sharing, and preservation of complex compound documents with embedded digital audio, video, and other dynamic data types. PDF permits communication of these published documents across all major types of computer platforms (MS-Windows, Apple Macintosh OS, and many flavors of UNIX, including Linux) and across time. Although it does not work for all types of data, it is good enough for archiving of many types of high-quality digital documents and data.

Extensions to PDF or other products still need to be developed to support permanent archiving of large commercial databases, spreadsheets, and other specialized scientific and technical data sets.

Adobe Systems is an example of a company with a truly unique philosophy and commitment regarding long-term product compatibility. Adobe has a broad and public commitment to provide 100% backward compatibility of the PDF format - for at least the next 25 years. No other software vendor has shown such a forward-looking philosophy and commitment to its customers. An important business reason for this commitment is that Adobe is beholden to the publishing industry, which has standardized on Adobe's PostScript language and many of Adobe's other multimedia publishing products. PDF is a new and improved but compatible version of PostScript that also supports high quality publishing on the World Wide Web. More than 140 government agencies worldwide have adopted PDF as their standard for document submission and archiving. Many industries have followed the lead of these government bodies and adopted PDF, primarily because it makes it easier to do business internally and with the government.

With extensions to support XML, PDF can apply to all types of data published on the web and elsewhere, and for exchange and archiving of more specialized scientific and technical data types. The most recent release of PDF added security features to support signing and witnessing of PDF documents, allowing users to create self-contained, portable electronic records. Finally, an army of third-party software developers is using the open and published PDF format to extend and apply PDF to different applications and niche markets.

*The XML Approach*

XML stands for Extensible Markup Language, and has been hailed as the 'silver bullet' for information storage and processing. XML is platform independent, which allows for easy transfer of information sets from one machine to another, without having to worry if the recipient of the information has the same software applications to open the document as the originator. This has positive implications for XML as a longer-term storage format.

Many people hope that the World Wide Web and technologies such as the XML will offer some stability and solve universal access problems. However at the current frenetic pace of innovation and market competition, chaos will remain the norm until enough end users and suppliers are organized and work together effectively to produce stability in the marketplace. Currently, nearly everyone using computers is forced into a defensive position of preserving records in the lowest common formats, in the worst cases, by printing to paper.

## Conclusion

Traditionally, digital systems were not designed for long-term preservation. For this reason the intellectual perspective of archivists and records managers needs to be transferred to the designers of digital systems in order to produce universal digital archiving systems and formats

Finally, while we must naturally be concerned with the quality and durability of digital storage media, and seek ever-better alternatives to the current generation of hard drives and optical disks, a still larger concern is the inevitable obsolescence of the systems needed to retrieve and read the preserved texts, and the paramount need for "backward compatibility" of systems. Simply stated, Information Technology of the early 21st Century and the organizations that use IT are generally ill prepared to prevent damage or loss of valuable electronic records or data.

## References

Arnita, J & Cantelon, P. 1993. Corporate archives and history: making the past work. Available: http://www.loc.gov/catdir/toc/becites/91-46918.refs.html (Accessed 12 May 2003).
Emberton, D. 2002. The digital dark age. Available: http://www.shift.com/print/web/385/1.html (Accessed 12 May 2003).
Kuny, T. 1997. A digital dark age. challenges in preservation of electronic information. Workshop: Audiovisual and Multimedia joint with Preservation and

Conservation, Information Technology, Library Buildings and Equipment, and the PAC Core programme. Available: *http://www.ifla.org/IV/ifla63/63kuny1.pdf* (Accessed 12 May 2003).
National Archives and Records Service South Africa. 2003. Homepage. Available: http://www.national.archives.gov.za/ (Accessed 10 May 2003).
Rothenberg, J. 1998. Avoiding technological quicksand: finding a viable technical foundation for digital preservation. Spiral Development & Evolutionary Acquisition. SEI Workshop September 13, 2000 Available: www.sei.cmu.edu/cbs/spiral2000/september/Rothenberg.pdf (Accessed 2 May 2003).

**Further readings**

Bearman, D. 1999. Reality or chimeras in the preservation of electronic records. *D-Lib Magazine* 5 (4). Available: www.dlib.org/dlib/april99/bearman/04bearman.html (Accessed 5 May 2003).
Bergeron, B. 2002. Why your digital data could one day disappear. Harvard Business School. Working Knowledge for business leaders Newsletter. Available: http://www.zacha.org/pipermail/cyberculture/Week-of-Mon-20020211/001025.html (Accessed 20 June 2003).
Bullock, T. 1989. Stone circles. Global community experience: 1989 – 2001. Available: http://www.globalcommunityx.com/europe/britishisles/stonecircles/history/histmenu.html (Accessed 15 June 2003).
Cullen, C. 2000. Authentication of digital objects: lessons from a historian's research. Available: http://www.clir.org/pubs/reports/pub92/cullen.html (Accessed 15 June 2003).
Diane, F. 2003. Electronic archiving enforcement lacking. *Federal Computer Week*. Available: www.fcw.com/fcw/articles/2003/0303/web-nara-03-04-03.asp (Accessed 3 June 2003).
Knoll, A. 2000. Digital Access to Old Manuscripts in the Memoriae Mundi: Series Bohemica Program. *Slavic & East European Information Resources* 3(2/3): 169-178.
Moore, R. 2001. Preserving electronic records for posterity. *Envision* 17(4). Available: http://www.npaci.edu/envision/v17.4/nara.html (Accessed 12 May 2003).
Rauber, A & Aschenbrenner, A. Part of our culture is born digital – on efforts to preserve it for future generations. Vienna, Austria. Available: www.ifs.tuwien.ac.at/~aola/publications/trans10.html (Accessed 4 May 2003).