Scientific software solutions

# ARCHIVING ELECTRONIC INFORMATION

**William Roberts**

**TESSELLA SUPPORT SERVICES PLC**

Issue V1.R2.M1
September 2004

# LONG-TERM PRESERVATION OF ELECTRONIC INFORMATION

## Introduction

Keeping records is vital for any organisation, and although such records potentially form a mine of information, it has traditionally been difficult to make full use of this resource owing to the relative inaccessibility of paper based records. As advances in technology increase apace, more and more of these records are now being generated, processed and stored in electronic format. This presents the opportunity for making the data readily available on-line and hence opening up a largely untapped resource to a much wider audience, potentially with substantial business benefits.

Before this can be put into practice, there are several issues associated with long term storage of electronic data that need to be addressed. With paper based records, the storage medium and format of written or printed information have remained essentially unchanged over hundreds of years. In contrast, both the format of electronic records and the medium on which they are stored can easily become unreadable within a few years. One only has to think of the number of now obsolete hardware and software formats that were industry-standards 15 years ago to begin to see the magnitude of the problem. The difficulties are exacerbated for organisations that are required by law not only to keep records for several decades, but also to demonstrate their authenticity.

This article examines the issues involved with archiving and preservation of electronic data, looks at the requirements for record keeping from a regulatory perspective, and examines some of the current approaches to the problem.

## Some definitions

The language of the record-keeping community uses some common words in specific ways, so it is worth stating the definitions of a couple of key terms. The following are taken from the reports of the InterPARES project [1].

> *Record: "any document made or received and set aside in the course of a practical activity".*
>
> *Authentic record: "a record that is what it purports to be and is free from tampering or corruption".*

Digital (or electronic) records, that is records stored on a computer system, whilst serving largely the same purpose, are inherently different from paper records. To access a paper record involves the user looking at and understanding the words, numbers or pictures on a piece of paper. To access a digital record requires the interaction of various items of hardware and software, combining to present the record to the user, typically on a computer screen. It is important to note that the record itself is this presentation or rendition of the information to the user, not the bit-stream of the computer files involved, or the section of the tape or disk where these bits are stored. Preservation of an authentic digital record means that we preserve the ability to present the information to the user, in such a way that it passes on the message intended by the creator of that record.

**Archiving vs. Backup**

A digital archive is a managed collection of digital records. It is worth emphasising the distinction between archiving and backup as in the computing literature 'archiving' is sometimes used in a less precise sense than we are using here. Backing up is a regular process required while data is changing on a regular basis. This process typically happens daily, with data retained for a period of months. Restoration is usually within a relatively short period of time, is typically requested by the author, and the prime method of accessing the data is by date. In contrast, the information within a digital archive needs to be accessible many years hence, perhaps by people who were not involved in any way with its generation. For this reason, other related information needs to be stored alongside the original data, and this is usually referred to as metadata - this is discussed in more detail below. Not only does this help to provide a context which makes the information easier to retrieve, it can also be used to store vital regulatory evidence as described next.

**Requirements for Digital Record Keeping**

Records provide evidence of a transaction or a state of affairs. As defined above, an authentic record should be demonstrably "what it purports to be and … free from tampering or corruption". Different groups have proposed different approaches to achieving this, and the requirements may differ in different contexts. Suppose for example that there was a legal dispute over the ownership of information stored in two separate computer files. It is all very well for one party to claim ownership on the basis that their file has an earlier date stamp for example, but this clearly does not constitute proof. An archive of digital information must therefore address the issue of evidence, if it is to be of any use from a legal perspective. The set of metadata associated with a record is an essential part of this: the metadata set should include contextual information, it should document the series of events from the record creation through various preservation actions to its current state and if necessary can involve a digital signature or other authentication technology. The use of digital signatures in recordkeeping is a key element of the US Food and Drug Administration (FDA) regulation 21CFR Part 11 [2], which governs how pharmaceutical companies must manage digital records to be included in submissions to the FDA.

A useful and comprehensive review of the issues surrounding digital recordkeeping, together with guidelines for best practice has been published by the UK National Archives (Public Record Office and Historical Monuments Commission). [3]. The US Deparment of Defense has also issued influential guidelines [4] including a list of metadata it requires to be stored as part of records.

**An explanation of the problem**

The technology required to store and render a digital record can be divided into three main components:

- the storage medium - e.g magnetic tape or disk
- computer hardware , including a device to read the storage medium (e.g. disk drive) and a processor to execute the programming instructions
- computer software, including operating system, device drivers and application software, to extract the stored information from the disk, process it and present it to the user in the required way

Storage media have a finite lifetime, so at regular intervals information on tape or disk must be copied to new media to avoid information loss. Because of the speed of technological developments in this area, the lifetime of the whole media format may be no longer than the lifetime of the media. In the last 15 years, the market leading magnetic tape format has probably gone through at least 5 or 6 incarnations. Therefore when it is time to renew the storage media holding the stored bit-streams, it is probably necessary to change to a new media format. Typically the new format offers higher capacity and quicker access, so the work required in the transfer is compensated to some extent by improved performance. Although this process needs to be carried out conscientiously, it is a common practice and presents no serious difficulties. Off-the-shelf storage management systems often incorporate automatic media checking and renewal facilities.

Similarly, computer hardware has a finite lifetime and when a computer needs to be replaced, not only does it usually make economic sense to buy the latest model, the manufacturers are usually unable or unwilling to supply any hardware with a specification more than a few months old. Again this is in itself a manageable problem, though care must be taken to maintain compatibility with storage devices and other peripherals.

The biggest obstacle to digital preservation is changes in software, both in terms of operating systems and application software. Access to the *information* contained in the record requires not only the stored bit-stream but also software to process those bits, in order to present them in the form intended. This is true of all file formats from ASCII through word processors to specialised CAD/CAM or database systems. Organisations frequently upgrade their software to new versions with new features. Keeping the original software associated with a particular stored file presents a number of difficulties: careful management is required to know which file goes with which software, maintaining multiple versions of software is a major headache for the IT department, manufacturers withdraw support for older products and often the killer blow is that older software will not run on new operating systems or new hardware, so that replacing computers necessitates replacing old software.

The challenge of digital preservation is thus to maintain the ability to present the original information in a record to the user, in the way the creator intended, using a complex network of hardware and software components, where every one of these components may need to be replaced several times over the lifetime of the record.

**Possible solutions**

Firstly, let us briefly list and explain the approaches that have been proposed to the long-term digital preservation problem. There is no widely accepted complete solution to this difficult problem, so the following section inevitably tends to focus on the drawbacks of each approach. This is not to say that there is nothing that can be done and the next section discusses practical approaches that can be taken now.

*Hard copy*

This refers to printing the document and storing the paper, or more sophisticated but essentially equivalent techniques using microfilm. We believe that this is not an acceptable or viable long term approach. By sacrificing machine readability, we lose many of the advantages of the electronic working environment, such as the ability to search, transmit, make perfect copies, etc. Also, any dynamic functionality of the record is lost and as digital systems diverge further from a kind of “electronic paper” then the ability to make a meaningful version of the record by printing will decrease.

*A Museum of Old Technology*

One possibility would be maintain old hardware and software in working condition, so that the original record can be rendered in its original environment. This is clearly only a short term solution since it would soon become prohibitively expensive to maintain. This approach has been used in practice by organisations needing a stop-gap measure until other solutions can be prepared.

*Emulation*

Emulation of old software on new hardware/software is another approach. Although this has been proposed by some commentators, notably Jeff Rothenberg [5] as a generic solution to the preservation problem, it is not widely used in practice. There is a significant amount of initial investment required in developing reliable emulators. It may turn out only to be feasible when vendors provide the facility as part of an updated or replacement product. In the long term, there is also the problem of needing to retain the ability for users to work with old software in an unfamiliar environment. This may be of historical interest perhaps, but is not ideal for the maintaining access to the information in general.

*Migration*

Migration in one form or another is the most widely applied approach at the moment. The term migration is used to describe many aspects of digital preservation actions, including copying of data from one media type to another: however the most common meaning relates to changes in the format of stored data files to make them accessible by new technology, e.g. a new operating system or new application software. An example would be migrating word processing files in Microsoft Word 95 format to Word 2000 format. In any migration process, there is the possibility that unwanted changes to the record are introduced so thorough testing and evaluation of any procedure is required before it is carried out, to ensure that any changes are acceptable from the point of view of record authenticity. The scale of the migration task grows as the archive gets larger and larger.

A thorough review of migration as a preservation strategy can be found at the Dutch government Digital Preservation Testbed web site[6]

*Standardisation*

This is a variety of the migration approach. Storing data in an industry standard format may help to reduce the rate of technical obsolescence and so reduce the required frequency of migration. A careful choice of formats may also make future migration processes easier and less error-prone. An objective of storing data in a standard format is to make use of well-documented and widely available standards to encapsulate the knowledge needed to understand the information in a digital file, rather than depending on proprietary software, where the file format and method of processing are less transparent.

The use of XML has been proposed for many aspects of standardised data formats, data interchange and software interoperability, and hence also has many possible applications in the field of digital preservation. The advantages of XML are that it is a well-structured, self-describing data format, easily processed by computer programs, yet reasonably easy for a human to read. It is ideal for structuring and storage of metadata. One of the most important advantages of this format is that it is a published open standard which is widely supported by all of the major computer software suppliers and so is likely to be a good choice for standardisation for a number of years to come. When it does become desirable to migrate from XML to some new approach, it should be relatively easy to do this in an automated and reliable way.

The Australian State of Victoria was one of the first public organisations to set up an operational digital preservation strategy. Their “Victorian Electronic Records Strategy” [7] takes a standardisation approach, specifying the required format of an electronic record. This involves an XML wrapper, containing the metadata, with the record content in PDF format, base64 encoded and encapsulated within the XML wrapper. Their approach also allows for the use of digital signatures.

XML allows the content and presentation of a document to be separated and this is widely regarded as a very positive feature of XML for data and document storage. One of the drawbacks at the moment for the use of an XML-only approach to record storage is that the XML based technologies for specifying appearance are still in a state of flux. However, the XML Stylesheet Language (XSL), including XSL Transformations (XSLT) and XSL Formatting Objects (XSL-FO) are promising in this regard and their use is becoming more widespread. XSL Version 2.0 is currently being drafted by the World Wide Web Consortium.

Developments based around XML are promising for use in digital preservation, but the short history of computing reminds us that we should not assume that any standard will last for very long, so standardisation is not a cure-all. Also, as Rothenberg [5] says, quoting Andrew S. Tanenbaum: “One of the great things about standards is that there are so many different ones to choose from!”

*Virtual machines*

This is an approach taken to make software programs interoperable between different types of hardware. Although it is not a new approach, it has become very popular in mainstream computing in recent years, with the Java platform and more recently with Microsoft’s C# and VB.NET languages. In this approach, the computer program is executed on a virtual machine - a piece of software which functions essentially as a type of translator that interprets the instructions of the computer program and translates them into something that can be executed on a particular real computer. For each new type of computing hardware, the virtual machine software only needs to be written once and then all programs written in the corresponding language can be executed on the new machine. This means for example that all Java programs can be executed on any machine for which a Java Virtual Machine is available - including more or less all varieties of Windows, Unix and Linux, as well as cut-down versions on less obvious devices such as mobile phones, PDAs and all kinds of consumer electronics.

Raymond Lorie of IBM has proposed an approach to long-term preservation based on this concept, which he calls the Universal Virtual Computer (UVC) [8]. He has defined a language and corresponding virtual machine involving a very simple instruction set and minimal assumptions about the architecture of the target hardware, so that even if computing architecture changes dramatically - and we can expect that it over the next century that it will indeed - it will still be straightforward to write new virtual machine software for each new platform. IBM has carried out a proof of concept for the Dutch national library (Koninklijke Bibliotheek) where a UVC program has been written to transform PDF files into a self-describing “logical data description”. The idea is that the UVC program and the original file are preserved. In a future computing environment, the UVC program is executed on the Universal Virtual Machine implementation for the new platform, to convert the PDF file to a readable Logical Data Description. Software can then be written for the new platform to process this file as required.

**Features of a preservation strategy**

This article has discussed two of the requirements of a digital archive, namely the maintenance of the evidential value of records, and the accessibility of content of those records over long periods. Off-the-shelf systems do not adequately address both these requirements at present. The majority of information systems are not designed to deal with preservation of the context of a business transaction, nor do they typically address the issues of long-term preservation.

Electronic document management systems (EDMS) and electronic record management systems (ERMS) go some way in this regard and are used by many organisations for example to record the review and approval of documents in order to satisfy regulatory requirements. They can assist greatly in the short to medium term management of records, including record classification and retention management, and can facilitate the automatic collection of metadata to capture information on the context of records. However, the majority of these systems are not well suited to handle the large volumes of laboratory or other technical data that some industry sectors must keep track of. Also, the great majority of currently available EDMS/ERMS have not been designed with long-term preservation in mind. In any case, it would be difficult for a single product to tackle the widely varying approaches to the organisation of data that are found even within a single industry sector. It is for these reasons that many organisations have been taking the approach of developing bespoke systems.

Before deciding on the type of system, it is of course necessary for an organisation to review the purpose for which it is keeping records, the kinds of facilities required for accessing the archive, and the retention periods of the record to be preserved.

As explained in Section 5, a system to preserve digital records involves a number of components, including storage media, hardware and software and in the long term, all components of the digital archive will need to be replaced, possibly several times (or in the case of government organisations planning to preserve some records in perpetuity, a potentially unlimited number of times).

In order for this to be successfully achieved, the architecture of the system must be highly modular, with the interfaces between the modules strictly defined. Communication between the modules should be in a format with the best chance of longevity, thus with present technology, an XML based messaging approach is a good choice. This allows individual components to be replaced without affecting other parts of the system. The multi-tier layered architecture common in modern business data management applications is well-suited to an archiving system, separating the data, the "business logic" i.e. the understanding of how to retrieve meaningful data from the data storage system, the presentation layer and the user interface.

The system must be protected from damage, whether it be due to hardware failure, physical damage such as fire or malicious digital attack, for example by viruses or hacking. Therefore the system must be duplicated or backed up in such a way that it can be completely and reliably rebuilt and it must be protected from unauthorised access. It should be incorporated in a disaster recovery plan It must include an audit trail so that the history of all changes to a record is preserved (e.g. migration to a new format, addition of metadata, renewal of digital signatures, etc.)

The degree to which standardisation is possible varies greatly from organisation to organisation, but in general it is better from the point of view of long-term preservation to choose as few different formats as possible and to choose widely used standards with publicly available specifications. The more readily available the specification of the data format, the better the chance that a file based on this format will still be readable far into the future. If a proper specification is available, new processing software can be written in the future if required.

Our message is that practical solutions to long-term digital preservation are possible. There are undoubtedly many difficulties, but it is not a viable strategy to do nothing, waiting for the perfect solution to arrive. The volume of digital records needing preservation is set to accelerate rapidly in the next few years. Organisations need to start now with their digital archives and to take the kind of future-proof flexible approach outlined above, so that not they are not permanently committed to a single strategy.

 

## References


[1] InterPARES (International Research on Permanent Authentic Records in Electronic Systems) www.interpares.org/

[2] US Food and Drug Administration. www.fda.gov/ora/compliance_ref/part11/

[3] UK Public Record Office, “Management, Appraisal and Preservation of Electronic Records”. www.nationalarchives.gov.uk/electronicrecords/advice/

[4] US Department of Defense http://jitc.fhu.disa.mil/recmgt/standards.htm

[5] Jeff Rothenberg, “Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation”. www.clir.org/pubs/reports/rothenberg/pub77.pdf

[6] Digital Preservation Testbed project www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=206&categorie=2

[7] Victorian Electronic Records Strategy. www.prov.vic.gov.au/vers/published/final.htm

[8] Raymond Lorie. “A project on preservation of digital data”. www.rlg.org/preserv/diginews/diginews5-3.html#feature2


 

**Tessella Support Services plc**

**Creating Software for Science and Engineering**

Tessella's services range from feasibility studies, through system design, development, implementation and ongoing support. Our expertise includes:

Data Analysis Software

Data Capture

Simulation Software

Advanced Graphics

Systems Support

Database Applications

## Other Technical Supplements available include:

- ❑ Active Server Pages
- ❑ Archiving of Electronic Info
- ❑ Automated GUI Testing
- ❑ Bayesian Statistics
- ❑ Beowulf Clusters
- ❑ C++
- ❑ Choosing and Using a LIMS
- ❑ COM
- ❑ Computational Fluid Dynamics
- ❑ Computer Image Processing
- ❑ Decision Support Systems
- ❑ Development for the Mobile Platform
- ❑ e-GIF
- ❑ Electronic Data Capture
- ❑ Electronic Lab Notebooks
- ❑ Evolutionary Computing
- ❑ Excel
- ❑ Extending the Life of Software
- ❑ FDA 21 CFR Part 11
- ❑ Formulation
- ❑ FORTRAN 90
- ❑ Grid Computing
- ❑ High Throughput Experimentation
- ❑ High Throughput Screening
- ❑ Instrumentation
- ❑ Integrated Lab Systems
- ❑ J2EE
- ❑ Java
- ❑ Linux
- ❑ Microsoft .NET
- ❑ Object Oriented Programming
- ❑ Open Source and Free Software
- ❑ Pocket PC
- ❑ Portable GUI Development
- ❑ Real Time Systems
- ❑ Regression Testing
- ❑ Security and the Internet
- ❑ Simulation
- ❑ Soft Computing
- ❑ Software Development Cycle
- ❑ Software Documentation
- ❑ Software Portability
- ❑ Software Re-engineering
- ❑ Software Specification
- ❑ SQL
- ❑ UNIX Inter-Process Comms
- ❑ UNIX Systems Performance
- ❑ Web Services
- ❑ Windows 2000 Services
- ❑ XML
- ❑ X Windows

**Tessella Support Services plc**

3 Vineyard Chambers, Abingdon, Oxon, OX14 3PX, England

Tel: (+44) (0) 1235 555511 Fax: (+44) (0) 1235 553301

E-mail: info@tessella.com Web Address: http://www.tessella.com