Long Term Data Storage: Are We Getting Closer to a Solution?

A. Stander & N. van der Merwe

Dept of Information Systems

University of Cape Town

And

astander / nvdmerwe@commerce.uct.ac.za

S.F. Rossouw

Codata Executive (South Africa)

steveros@iafrica.com

Abstract

Many scientific and socioeconomic reasons exist for the long term retention of scientific and lately also business data. To do so successfully, the solution must be affordable and also technologically flexible enough to survive the many technology changes during its useful life.

This paper looks at the current status of available technology for long term data storage, more specific the standards that exist for data interchange, the creation and storage of metadata, data conversion problems and the reliability and suitability of digital storage media. Even if in the ideal format, application and database management software is needed to store and retrieve the data. Typically the life expectancy of such software is much shorter than that of the storage media and as this has already been the cause of major data loss, possible solutions are investigated.

Most research into long term data storage focus on large to very large databases. It is often forgotten that small, but very important pockets of scientific data exist on the computers of individual researchers or smaller institutions. As most of the time this is stored in application specific formats with a short lifespan, strategies for the preservation of smaller amounts of data are also looked at.

Introduction

Preserving electronically-held data is a complex and challenging problem. The instability of the storage media and the rapid obsolescence of the equipment needed to read the data pose problems which can seem insoluble in the face of accelerating technological change. For archival purposes it is hoped that optical discs, with a lifetime of up to 30 years, will offer a solution. This is however only a partial solution as the hardware and software needed to read these products is usually unsupported within a decade. (Mackenzie, 1991).

Documents which are delivered on a physical medium present problems in respect of the substrate itself. With magnetic media, the signal on the substrate tends to attenuate fairly rapidly, and it is customary to rewrite records on magnetic tape on a regular basis, often annually, in order to preserve them uncorrupted. Magnetic media are also vulnerable to corruption by magnetic and electrostatic fields and physical and chemical changes to the plastic substrate caused by incorrect storage.

Archival Format

A possible solution to many of these problems is stripping away software and machine dependence from the data, documenting it to preserve provenance and meaning, and converting it to a common character format ASCII or Unicode, which will then require periodic refreshment by re-copying to the most stable medium currently available (Ross & Higgs, 1993).

Unicode provides a unique code for every character, no matter what the platform, program or language. The Unicode Standard has been adopted by many industry leaders and is required by modern standards such as XML, Java, JavaScript, LDAP, CORBA , WML, etc., and is the official way to implement ISO/IEC 10646. It is supported in many operating systems, all modern browsers, and many other products. The emergence of the Unicode Standard, and the availability of tools supporting it, is a significant development in global software technology.

Incorporating Unicode into client-server or multi-tiered applications and websites offers significant cost savings over the use of legacy character sets. It allows data to be transported through many different systems without corruption (Ross & Higgs, 1993).

The costs associated with the need to recopy data are high, and there may be problems of copyright. Moreover, whilst such techniques may be feasible for straightforward textual and statistical data, in many cases severing the link between the software and the records renders the records themselves meaningless.

For documents which are delivered over networks, and which may have no physical existence other than as files on a hard disc somewhere in the world, it is necessary to ensure that they continue to be available. This implies the development of a set of standards for online publication, or the development of a structure, at national or regional levels, of archives for electronic materials.

Physical decay of media and technological obsolescence cover only a minor part of the difficulties in long-term data preservation. Preserving the pure data, does not guarantee that the information it holds is preserved along. Software is necessary for making the digital documents accessible and meaningful. Programs are supplemented by others that offer greater functionality and data formats in a single application are often adapted in order to achieve this.

New versions of the same software are often not similar enough to its predecessor to ensure that no information is lost on conversion. Exacerbating this is the dependency of application software on the underlying operating system, which in turn requires a specific type of hardware platform (Russell,1999). As a consequence, a future system will most probably not be able to run the program on a later hardware version.

SGML & XML

SGML is the Standard Generalized Markup Language (ISO 8879:1985). It is an international standard for the definition of device-independent, system-independent methods of representing texts in electronic form (Goldfarb, 1990).

SGML is very large, powerful, and complex. It has been in heavy industrial and commercial use for over a decade, and there is a significant body of expertise and software to go with it. XML is a lightweight cut-down version of SGML which keeps enough of its functionality to make it useful but removes all the optional features which make SGML too complex for some applications.

XML is intended to make SGML easy to use. them across the Web. For this reason, XML has been designed for ease of implementation, and for interoperability with both SGML and HTML. It is not just for Web pages: it can be used to store any kind of structured information, and to enclose or encapsulate information in order to pass it between different computing systems which would otherwise be unable to communicate (W3C, 1998)

SGML is a metalanguage, that is, a means of formally describing a language, in this case, a markup language. Historically, the word markup has been used to describe annotation or other marks within a text intended to instruct how a particular passage should be printed or laid out. As the formatting and printing of texts was automated, the term was extended to cover all sorts of special markup codes inserted into electronic texts to govern formatting, printing, or other processing.

Descriptive Markup

A descriptive markup system uses markup codes which simply provide names to categorize parts of a document. Markup codes such as <para> or \end{list} simply identify a portion of a document and assert of it that “the following item is a paragraph," or “this is the end of the most recently begun list," etc (Goldfarb, 1990).

A procedural markup system defines what processing is to be carried out at particular points in a document: In SGML, the instructions needed to process a document for some particular purpose (for example, to format it) is sharply distinguished from the descriptive markup which occurs within the document. Usually, they are collected outside the document in separate procedures or programs.

With descriptive markup the same document can readily be processed by many different pieces of software, each of which can apply different processing instructions to the relevant parts of it which are considered relevant. For example, a content analysis program might disregard entirely sections needed by a formatting program. Different sorts of processing instructions can be associated with the same parts of the file. For example, one program might extract names of persons and places from a document to create a database, while another, operating on the same text, might print the text in a specific format.

Types of Document

SGML also introduces the notion of a document type, and hence a document type definition (DTD). Documents are regarded as having types, just as other objects processed by computers do. The type of a document is formally defined by its constituent parts and their structure (Goldfarb, 1990).

If documents are of known types, a special purpose program (called a parser) can be used to process a document and check that all the elements required for that document type are indeed present and correctly ordered. Different documents of the same type can be processed in a uniform way. Programs can be written which take advantage of the knowledge encapsulated in the document structure.

Data Independence

A basic design goal of SGML was to ensure that documents encoded according to its standards should be transportable from one hardware and software environment to another without loss of information.

Descriptive markup and document typing both address this requirement at an abstract level; the third feature addresses it at the character of which documents are composed. SGML provides a general purpose mechanism for string substitution, that is, a simple machine-independent way of stating that a particular string of characters in the document should be replaced by some other string when the document is processed (Goldfarb,1990)

One obvious application for this is to counter the inability of different computer systems to understand each other's character sets, or of any one system to provide all the graphic characters needed for a particular application, by providing descriptive mappings for non-portable characters..

MEDIA

Appropriate selection, storage and handling of media is essential to any preservation strategy. It is important to have an understanding of the various media for storage because they require different software and hardware for access, and have different storage and preservation requirements. They also have varying suitability according to the storage capacity required, and preservation or access needed.

Obsolescence of storage media has occurred in rapid succession. In floppy disks alone we have seen a progression from 8 in to 5.25 in and then 3.5 in formats, with each change leading to rapid discontinuation of previous formats and difficulty in obtaining or maintaining access devices for them.

Magnetic Media

Magnetic consist of a variety of magnetic media and containers including a range of magnetic tapes such as reels, cartridges, cassettes and disks. They all utilise the magnetic properties of metallic materials suspended in a non-magnetic mixture on a substrate.

This provides a flexible, low cost storage medium and both the storage capacity and the ability to retain the magnetic charges have increased substantially in recent years. The method of construction and storing the data also point to potential weaknesses of magnetic media.

Strong magnetic fields may alter the media and lead to data loss but this is rare and the media normally has to be in very close proximity for this to occur.

Clean operating conditions and environments will reduce the scope for damage to media and devices. The high density of storage and the close proximity of device heads to the media mean even small particles such as smoke or other debris can lead to

data loss.

Poor environmental storage may also lead to oxidation of the ferromagnetic material or problems with the "binding" layer or substrate materials.

Handling and use of magnetic storage media should be minimised to reduce wear, or refreshment cycles implemented to replace media on a frequent basis reflecting the levels of use.

It is also important that attention is paid to the recording and access devices. These should be of good quality and well-maintained. Problems with the access devices e.g. head/media crashes are one of the most common cases of damage to magnetic storage media.

Optical Media

Optical storage media use laser light to read from a data layer. A number of different types such as CD-ROM (Compact Disc - Read Only Memory), CD-R (Compact Disc -Recordable), and DVD-ROM (Digital Versatile Disc - Read Only Memory) exist.

In CD-ROM the data layer consists of a series of pits in a metallic coating over a plastic disk. A clear acrylic coating is applied to the metallic layer to protect it from scratches and corrosion. CD-R employs a dye layer which is light sensitive as the data layer. The use of light sensitive dyes means CD-Rs are less stable than CD-ROMs as archival media (Ross & Gow, 1999).

As with magnetic media there is considerable diversity in practice and production of CD-R and care is needed in selecting high quality media. DVD-ROM is a more recent optical storage medium with capacity to store up to 18 Gb.

Optical disks are an increasingly popular method of storage. The device reader is not in contact with the disk and mechanical failure is less likely to lead to data loss than damage to the disk itself through poor handling or storage.

Media life

Media should be refreshed on a regular cycle within the lifetime for archival storage identified by the manufacturer or independent sources. Sample generic figures for lifetimes are given below (Ross & Gow, 1999).

- D3 Magnetic tape: 1 – 50 years
- DLT magnetic tape cartridge: 1 – 75 years
- CD/DVD: 2 – 75 years
- CD-ROM: 3 months – 30 years

Smaller Amounts of Data

Most principles applicable to large scale data preservation also apply to the preservation of smaller amounts of data.

Where possible, the most sensible solution seems to store smaller amounts of data as XML on CD-ROM using ISO 9660, Interchange Level 1 formatting. Using CD-R is easier to create for small amounts of data, but necessitates the implementation of a refreshment cycle to prevent data loss. This cycle can however be much longer than for magnetic media. Suitable storage facilities are an obvious extension of a successful archival strategy.

Without the necessary policies to make the data available to future users, it is obvious that no amount of preservation is worthwhile.

Management

Digital preservation raises many organisational, legislative and economic concerns. Yet, these issues can hardly be addresses in the absence of a sound, technical solution to the digital longevity problem (Rothenberg, 1999).

For systems which display inter-dependence between software and data, standards need to be developed for the taking of meaningful "snapshots" for archival purposes, and system software developers need some incentive to meet such standards (Morelli, 1993).

Similar standards are required in respect of systems where the data held is transient. These include both electronic publications distributed by non-material means (networks and broadcasting), and also most office systems.

A third problem affecting the long-term usability of

digital records is the failure to develop a migration strategy for moving records to new media and technologies as older ones are displaced. The unavoidable fact is that digital records are technology dependent and therefore outdated technology is likely to be the most serious impediment to the long-term usability of digital records. The development and implementation of a migration strategy to ensure that digital records created today can be both processed by computers and intelligible to humans in the future is absolutely essential.

Conclusion

It is clear that the sensible application recent developments can go a long way to solve some of the most critical problems associated with long term data storage.

The use of a hardware and software independent data format is facilitated by the wide acceptance of the XML format and the latest optical storage can easily offer secure storage of thirty years or longer if properly stored.

All problems are not solved however and strategies and standards are needed to handle transient data and the migration of data to new media. Sheer volume can often be a problem any the issue of copyright must also be addressed.

References

Farley, J. (1999). An Introduction to Archival Materials; new media (PRO Preservation Guide series).

Goldfarb, C. (1990). The SGML Handbook, Oxford.

MacKenzie G P (1991). Preservation of computer records. In *Information 90*, edited by Jennifer Rowley, (London, 1991), pp.343-347.

Morelli, J.D. (1993) *'Defining Electronic Records: a Terminology Problem ... or Something More'*, in Ross and Higgs (eds, 1993).

Ross S & Higgs E (1993). *Electronic Information Resources and Historians: European Perspectives*. British Library Report 6122. Gottingen: Max Planck Institut fur Geschichte.

Ross, S. & Gow,A. (1999). Digital Archaeology: Rescuing Neglected and Damaged Data Resources. British Library Research and Innovation Report 108. London,The British Library.

Rothenberg J. (1999). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library and Information Resources, January 1999 at http://www.clir.org/pubs/reports/rothenberg/contents.html.

Russell K. (1999). Digital preservation: Ensuring access to digital materials into the future. June 1999 at http://www.leeds.ac.uk/cedars/Chapter.htm.

W3C (1998). Extensible Markup Language (XML), Version 1.0