August 08, 2001

InSide Gartner (IGG)
D. Logan, K. Weilerstein, A. Weintraub

Management Update: Important Issues About Digital Data Preservation

Long-term preservation of digital data is becoming an issue for an increasing number of enterprises. Gartner discusses some of the most important aspects.

Many enterprise executives recognize the challenges of long-term digital data storage. For example, many IT managers are wrestling with the problem of the storage, maintenance and accessibility of digital data over the long term — a decade or more. Long-term preservation of digital data is becoming an issue for an increasing number of enterprises. To help enterprises with their planning, Gartner discusses some important issues about digital data preservation.

Relative Time Frames for Content Retention

What are the definitions of short, medium and long term? Retention periods for records are determined by legal and business requirements. The length of time a record must be kept and the frequency with which it is accessed will determine how and where it should be stored.

- **Short-term storage —** two years or less. For frequent (daily) access.
- **Medium-term storage —** between two and 10 years. Mandated in many countries for tax records. The records may be accessed only weekly, monthly or quarterly.
- **Long-term storage —** 10 years or more. Legally, records may have to be maintained for the life of a product, customer relationship or the life of the company. To support historical research, governmental agencies may keep documents in perpetuity. Individual records may be accessed extremely rarely, although the archive as a whole may be accessed daily. Each and every record (or group of records) in an archive should have an associated retention period.

Storage Media Strategies for Short, Medium and Long Term

Short-term storage should be on easily accessed magnetic media. Slower optical or tape jukeboxes, frequently termed near-line storage, are recommended for medium-term storage. Long-term storage may be offline, stored on analog media (paper or microfilm/fiche) or optical disk platters. Retrieval of this content can take days or weeks. These storage media get progressively less expensive per megabyte of information stored, excluding conversion labor and hardware costs.

Digital data is ephemeral. Migration to analog media circumvents the often-overlooked but very real problems with digital media: it decays and the access mechanisms become obsolete. Long-term data migration costs can be considerable and will result in the loss of some data. No one understands how to archive digital documents in a way that will guarantee their integrity and accessibility over the long term. Other than storage on traditional media, sustainable solutions to the problem are not currently available.

Longevity of Various Media, Refreshing Requirements
Vendors may claim that the current generation of 5.25-inch and 12-inch optical disk media and "top-notch" archival quality CD-R (compact disc-recordable) media will last 100 years or more when properly housed. That claim is in dispute, however, and Gartner recommends that optical media be tested periodically to determine whether degradation is beginning to take place.

Lower-grade media is known to last considerably less time, but no one knows for sure how long the media will really last, even under ideal circumstances. Records may, therefore, have to be moved to a new storage media every five to 10 years to keep up with technology and avoid degradation. Bear in mind that studies show that each migration may lose up to 5 percent of the stored information.

If archival media are in danger of deteriorating because of the effects of time, wear or because of poor preservation, they can be transcribed (refreshed) onto a blank disk or tape. Refreshing the media will reset the clock on deterioration, but it is no substitute for the migration to up-to-date media and computer systems. Freshly written disks or cartridges are only useful if there are drives to read them and computer systems to interpret what is read. Users who build their preservation strategy around refreshing media must also ensure that blank media and access devices will be available in the future.

Taking into account the frequency of access, retention period and volume of data, Gartner recommends a move to analog, human-readable media for records that are to be housed longer than 10 years. Microfilm has an estimated life of up to 500 years, if properly stored.

Any preservation strategy must be coupled with a records retention policy and records retention software (see the sidebar, "The Role of Records Management Software").

Media Longevity Determination
Proper storage in a controlled environment will ensure that any storage media will last longer than media in an office environment or warehouse. Follow the manufacturers' advice. Just as with music CDs and tapes, storage media have a finite number of read and write cycles; constant access will gradually obliterate the recording.

Write-Once Requirements
Although you may not wish or need to make any changes to archival material, in practice, imposing a write-once requirement on a digital archive increases costs, lowers performance and limits choice of storage technologies. Check whether write-once storage is mandated by law, or whether it is an internal requirement that may be negotiable. Laws and regulations change from time to time, and if there is any doubt, seek legal counsel.

The strictest definition of write-once means media that physically do not allow overwriting: CD-R media, the phase-change optical disk technologies made by Pioneer, Plasmon and IBM. Microfilm, microfiche and paper are legally considered as write-once media. This is unlikely to change.

A more-lenient definition may allow rewritable media that are protected by the cartridge and drive design from overwriting. This is available at a wider range of costs, capacities and speeds in the form of tape, CD, DVD (digital videodisc) and 5.25-inch optical media. Keep in mind what write-once protection cannot accomplish:

- It cannot prevent tampering before the document is archived or after it is retrieved.
- It does not do away with the need for off-site backups.
- Despite the confusing sales talk about “permanent” storage, write-once capability has no bearing on the longevity of the media.

Legal Requirements for Electronic Record Keeping

The legal requirements for electronic record keeping evolve constantly between jurisdictions and industries. There is no general advice. Gartner recommends monitoring legislative and regulatory activity via one of the recognized professional associations or governmental bodies (see the sidebar, “Resources for Legal and Regulatory Requirements”).

Formats for Text, Graphics, Images and Video

Text: PDF (Portable Document Format) is the most-prevalent storage format for purely electronic documents. PDF is compact and the viewer is free and widely available. The format is proprietary and there are concerns that PDF may not be recognized as a legally acceptable repository format because of its proprietary nature.

Graphics: JPEG (Joint Photographic Experts Group) is the standard for photographic or picture bitmap data. It is compact and widely used on the Internet to make picture files easy to download. Some data is lost using JPEG compression, so text is often hard to read in JPEG files. New techniques include JPEG2000 (a lossless compression technique), JBIG2 (Joint Bitonal Image Group 2) and MRC (Mixed Raster Content).

Images: TIFF (Tagged Image File Format) is the most widely used format for image data, especially when many documents are scanned daily or to archive them. It is largely platform-independent, making it readable and processable on a variety of hardware and software platforms; however, there are variations in the TIFF format, which may cause compatibility problems. If TIFF files are compressed they are less portable than uncompressed files. Uncompressed TIFF is the recommended archival standard for scanned images, despite their size.

Video: The MPEG-4 (Motion Picture Experts Group-4) standard for digital video is another example of a nonstandard “standard.” Vendor rivalry is precluding any widespread adoption of the standard.

Many organizations have been storing SGML/HTML and are beginning to store XML.

As continuous technology improvements drive business decisions, formats that have been considered de facto standards will be superceded. Obsolescence will be a factor in format selection, just as it is in hardware and software selection. Format migration will undoubtedly be necessary, no matter what standard is selected. This can be seen in the proliferation of standards in XML and graphics compression. Similarly, MPEG-4 has succeeded three previous standards.

Direction Being Driven by Government and University Archivists

No one — government agency, university, vendor or Gartner — has the ultimate answer to long-term preservation of digital records.

There are a number of trends.

- There is no short-term or long-term intention of abandoning "traditional media," paper, microfilm and microfiche.
- There is a trend is toward more and more digitization of records for greater accessibility. This will also drive down the unit cost of document access, which is labor-intensive.
- If forced to make a choice between converting paper records to microfilm or into digitized images, archivists recommend filming first.

Archivists as a group recognize the long-term costs of data migration and the issues surrounding data loss, both of which are significant. Like many enterprises, they are committed by default to data, software and hardware migration for digital data that is not otherwise available, as they have no alternative plans in place.

A second common theme among professional records management groups is the creation of methods to allow long-term preservation of digital media, especially those "born digital." Here, there are two primary areas of research: creation of a standard for preservation metadata, which will allow future generations to access information about how to access digital data (see the sidebar, "Preservation Metadata: Open Archival Information Standards") and technology emulation, reproducing older generations of technology using the current generation.

Bottom Line

- Digital data preservation problems need a strategy, not a product. The problems can't be solved by any single vendor.
- It is impossible to predict what the hardware and software of five or 10 years from now will be like, let alone the situation in 50 or 100 years.
- Because of the complexity and time scale of the issues, enterprises are increasingly turning to outsourcers to manage long-term preservation of data. Through 2005, 50 percent of nongovernmental enterprises with long-term preservation needs will turn to outsourcers (0.9 probability).

The Role of Records

Management Software

Records management helps enterprises control the volume of digital data that they store. It seems obvious that different types of records have different life spans, but many enterprises ignore this fact and tend to keep everything "forever." Records management evaluates documents for their fiscal, legal, operational and historic value, recognizing the need to minimize risk by periodically destroying items no longer needed. More than a space-saving measure, records management can help with decisions in designing and using electronic document repositories. Retention is systematic, ensuring that each record is kept as long as necessary, then systematically destroyed unless it is deemed to have historic significance or value in perpetuity. There is no reason records management products cannot be configured to help enterprises deal with records that have extremely long retention periods. Record retention policies and software do not solve the digital preservation problem, but they do help ensure that the only records that are kept are the ones that should be kept. Systematic records retention policies make digital data preservation a more-manageable proposition.

Preservation Metadata: Open Archival Information Standards

Metadata is a crucial component of a digital data preservation strategy. Emulation is dependent on it. At a minimum, preservation metadata must include descriptions of the hardware, software, formats and any access control mechanisms that were stored with the data. The Open Archival Standards group is working on an XML document-type definition for the construction and interpretation of this metadata. The Open Archival Systems reference model is a reference tool that describes a framework for defining a digital archive.

The techniques of technology preservation and technology emulation allow data to be stored as a bytestream, which is made accessible as needed. To access the data, formatting information is needed and this will likely be XML-based. However, XML should not be viewed as the final answer — in 100 years, all current formats will likely be obsolete and new ones will have taken their place.

What is needed is a way to preserve information about what our formats were and how they worked, so that historians in the future can access the data. Encapsulation is an important notion in this context. Encapsulation groups the data objects and any information necessary to provide access to them in one container or wrapper. The Open Archival Information System (OAIS) describes the metadata that needs to be there to facilitate encapsulation. Some national projects that are working with the OAIS reference model are the Cedars Project in the United Kingdom, the NEDLIB and the Preserving and Accessing Networked Documentary Resources of Australia (PANDORA) Project.

Resources for Legal and Regulatory Requirements

The Association for Information Management Professionals (www.arma.org) is an international body that can provide help and assistance to records management professionals. It has special committees that monitor government initiatives in the United States and Canada. The National Archives and Records Administration (www.nara.gov) runs a good site for general information on regulations and legal issues in the United States. In the United Kingdom, the most comprehensive government site is run by the Public Records Office (www.pro.gov.uk). The National Library of Australia has an excellent site that has resources for all aspects of digital preservation (www.nla.gov.au/padi).

The British Library, the Networked European Deposit Library (NEDLIB), the International Council on Archives, the International Records Management Trust and the European Commission on Preservation and Access (EPCA), along with a number of U.S. state government and university library sites, run excellent reference sites for legal and other digital preservation topics.

Written by Edward Younker, Research Products

Analytical sources: Debra Logan, Ken Weilerstein and Alan Weintraub, Intranets & Electronic Workplace

For related articles published in Inside Gartner This Week, see:

- **"Management Issues:** Long-Term Strategies for Digital Data Preservation," 4 July 2001

IGG: IGG-08082001-04

Entire contents, Copyright (C) 2001 Gartner Group, Inc. All rights reserved. Reproduction of this publication in any form without prior written permission is forbidden. The information contained herein has been obtained from sources believed to be reliable. Gartner Group disclaims all warranties as to the accuracy, completeness or adequacy of such information. Gartner Group shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.