

# Archiving Electronic Files

---

## IMERGE Consulting, Inc

---

*Bernard Chester*

*May/June 2006*

*Bernard.chester@imergeconsult.com*

*As Appeared in e-Doc Magazine*

*www.imergeconsult.com*

Maintaining long-term access to content can be difficult. Here are some things to consider as you archive your content.

Houston, there is a problem. The advent of electronic record producing and content management systems has created a new challenge, How do you keep electronic information long-term? Before the electronic age keeping records was simple. You boxed up the paper in an organized scheme, usually matching your filing plan, and stuck it in a warehouse. If the financial computations justified it, you might film the paper and store that in the warehouse. Barring disaster, you were sure that you would be able to read the information for decades. Electronic information is not so easily handled.

Electronic information such as email messages arrive in quantities that overwhelm manual approaches. In a large organization, manually sorting out those messages with information that must be kept would require a major organizational effort; printing them would consume forests and require warehouses. The performance of many application systems declines once they have filled with inactive information. Finally, it is clearly impractical to expect to keep a legacy software system running for years after it has been replaced. In most organizations, software and hardware systems are all amortized in no more than five years, and rarely are they kept in use more than 10 years.

Records and information managers have struggled with devising schemes to preserve important information for the long-term. Several solutions have gained currency:

Turn it into a physical record. Paper is something record managers have extensive experience managing. One option has been to print out the information, using archival quality paper so it doesn't turn brown or get brittle quickly, and store the records in the record center. The problem, of course, is that this can create an overwhelming quantity of paper, potentially unmanageable due to the manual effort required to support it. Even when this approach is feasible, access to an item of information can be expensive. Microfilm is another proven approach for keeping records long-term. Computer output to microform devices permit generating preservation quality film at a comparatively low cost since there is little need for manual involvement.

Transition the information to the new system. When transitioning from one electronic system to another, it may be possible to convert the contents of the old system into the new system. This assumes that all information can be completely and accurately moved, and the legacy data will not unduly burden the new system. To accomplish this may involve producing an export of the contents of the old product that can be imported into the new system, or pre-loading database tables in the new application. The transferred data must be quality assured to ensure that errors have not been introduced.

Convert to a standard format and store in an archive. Increasingly, archival electronic records are being converted to some standard non-proprietary electronic format (if not already in one), and stored and managed with technology. This can permit easy organization, low cost for storage, and the possibility of easy access to the contents if needed. Conversion is the approach we will focus on for the balance of this article.

Whatever approach is taken, the process must be carefully documented and recorded so future questions about accuracy and completeness may be answered. When planning to manage your long-term records using conversion, there are several issues that must be addressed: choice of format, structural documentation, media, and long-term management scheme.

Converted records are best kept in an electronic format that is non-proprietary, publicly documented, and actively supported. This will reduce the risk of later being unable to read the files. Even if the format is very popular, complete documentation should be acquired or developed, and archived with the records. The history of technology

shows that improvements and market forces can obsolete a technology within a few years. There are a large range of formats to choose from; the most popular ones group into several categories:

#### Text:

Pure Text: Alphanumeric information might be kept as text files. However, if choosing this approach, care must be taken to select and record the character encoding used. If multiple languages may need to be saved, then Unicode (ISO/IEC 10646) should be considered for the encoding, since it has a single value for every character, no matter the language.

XML: Extensible Markup Language has the advantage of providing structure and identification to the fields of data, making it easy to extract it later.

#### Still Images:

TIFF: Tagged Image File Format is a de facto standard for storing raster images. Most document scanning systems produce TIFF as their output. TIFF provides support for a number of color spaces, densities, compression methods, and pixel formats. Few viewers can properly handle all variants, so care must be taken to choose a combination that is popular.

#### Mixed Text and Picture formats:

HTML: HyperText Markup Language, the language of Web pages may seem appropriate when recording Web information, but has some challenges, as pages often have linked components.

PDF & PDF/A: Adobe's Portable Document Format has become a de facto standard for distributing documents, since it attempts to ensure a consistent appearance across viewing environments. PDF/A(rchive) is subset that has been made an international standard specifically for use in archiving documents.

JPEG2000 & DjVu: These formats are designed specifically to optimize the storage of documents with a mix of text and graphics.

#### Desktop Tool Formats:

Native Formats: Since many records are created using computer desktop tools such as Microsoft Office, drawing packages, CAD, and other applications, it is easiest to maintain them in that format. It provides the benefit of someone being able to use the file later to produce new records. However, these formats are generally proprietary and subject to vendor changes from release to release.

Open Document Format (ODF) and OpenXML have been introduced in recent months to move desktop applications towards open formats. Both of these proposals use XML to package and self-describe application information. Neither has yet achieved international standards body acceptance nor wide-spread availability, but in times to come, this may provide the ideal solution for desktop-generated records.

#### Audio and Video:

Of course one can always store audio and video with an analog approach. However, increasingly, with the advent of CDs and DVDs, digital techniques are becoming prominent. There exist a number of digital audio/visual formats, including WAVE, AIIF, MP3, Real Media, MPEG, AVI, and QuickTime. Many of the players support multiple formats, but it is difficult to be sure which will be the leader in 10 years.

With so many choices, how do you decide? Obviously, the type of information will be a big factor. But that may still leave one with a number of choices. My recommendation is to rate your choices favoring those that are least proprietary and most popular. This should give you the best chance of being able to access the record in the future. Whatever format is used, it is important to periodically check that your viewing tools still support it.

It will not be enough to have a library of accessible records in 20 years; you must also be able to interpret the data contained in the documents. The structure and interpretation of your information only you can provide. For example, the data fields in a billing record or report that has been saved need to be documented and explained to prevent later confusion.

The electronic media used to hold the archives must be reliable over time. Whether you choose a magnetic or an optical technology, media may fail or its underlying technology may become obsolete. Reliability may be achieved through duplicate copies or backups, fault tolerant storage systems, or the use of archival quality media.

When you build your archive, you will need some tools to manage the collection of electronic records. While a hierarchy of directories is simple and inexpensive, it will limit your ability to locate information, as well as to address records management issues. An electronic recordkeeping solution will provide more functionality, but it will require you to collect metadata that describes the record, documents its origins, and records management information.

After you have constructed your archive, regular testing and maintenance will be required, to ensure that you can still access the information and nothing has been lost. It might even be necessary to convert the archive to new technologies in order to continue its utility. Archiving electronic information requires some thought. If well planned and designed, you will be successful at protecting your organizations information.

Bernard Chester(bchester@imergeconsult.com) is a consultant on ECM who focuses on implementation and integration issues. He is a principal with IMERGE Consulting ([www.imergeconsult.com](http://www.imergeconsult.com)), an independent ECM consulting firm.