....

National Information Resources Face a "Phantom Menace"

www.mullermedia.com

> *A slice of America's history has become as unreadable as Egyptian hieroglyphics before the discovery of the Rosetta stone. Vast untold volumes of historic, scientific and business data are in danger of dissolving into a meaningless jumble of letters, numbers and computer symbols. Much information from the last 30 years is stranded on computer tape from primitive or discarded systems-unintelligible or soon to be so…*

The quotation is from a January 1991 Associated Press article[1]. As we approach the millennium a good deal of progress has been made in some quarters, but the scope of the problem is growing at an alarming geometric pace.

Who or what are the culprits? They're as stealthy and relentless as the most malicious hacker. They're making millions of computer files unavailable for further use—files that cost government and business billions of dollars to create. In a very real sense, *TIME* and its minions are the enemy: the "Phantom Menace".

- ♦ *TIME* and fragile storage media can cause valuable records to "decay" on the shelf.
- ♦ *TIME* and the inevitable migration to new computers, software and media renders older electronic records incompatible and unusable on the new systems.
- ♦ *TIME*, down-sizing and job-hopping programmers lead to undocumented programs and files that are difficult to decipher.

FACT: A greater portion of our national information resources is threatened by these factors than by terrorists and hackers.

A Variety of Reasons to Keep "Old" Data

Legal Mandates are the overriding consideration for government agencies and banks, among others. Concern over preservation of electronic records has reached new heights as court rulings emphasize the government's responsibility to retain e-mail and other computer files. The problems involved in preserving these records have also received increased congressional interest. The battle to protect these assets is being fought by public sector agencies such as the Center for Electronic Records at the National Archives and Records Administration (NARA). Financial institutions are an example of entities that are required by law to retain at least the last seven years' data available for inspection.

Historical Research is another motivation. Academics and journalists often access public electronic records in the course of their work. Federal, state and local governments provide a valuable (sometimes fee-based) service in preserving and providing access to such records. Business analysts need long-term information to evaluate company or industry trends. In some cases, information accumulated by a company as a by-product of its normal work may be of sufficient value to sell to the public.

Electronic Evidence. In more and more lawsuits, companies may need to deliver historical computer data as part of the discovery process. In criminal cases, disks, tape and hard-drives are often confiscated.

[1] Reading this particular article led to the author's involvement in related work at the National Archives.

Not infrequently, the data is in a form that's not readily usable by the court or other parties to the litigation.

The Freedom of Information Act (FOIA) has greatly expanded the government's responsibilities for and consciousness of preservation and access issues. FOIA cases can involve all three of the foregoing needs/motivations.

Obstacles to Preservation and Future Access.

The computer archivist must address both the future and the past. As for electronic records yet to be created, there is at least the hope that standards may be enacted that will ameliorate the difficulties of preserving and organizing that future ocean of data. The main trick will be to come up with standards that do not have the effect of stifling innovation. However, that complex arena is the subject for another article.

The focus of this article is on handling the vast amounts of existing non-standardized material. This mess needs to be cleaned up before we can proceed with the (hopefully more orderly) records of the future. Let's review the underlying considerations that confront both the computer archivist and the media conversion specialist. It's a bit like peeling back the layers of an onion:

- **Media**. The physical medium itself is the most obvious factor. Because the great bulk of accumulated information is on tape, we'll focus primarily that class of media[2]. In some instances the type of tape drive has become largely obsolete (e.g.- nine and seven-track reel-to-reel). In other cases there has been a fairly orderly progression of capacity within the same general medium (e.g.- 8mm helical scan tape and 4mm DAT.) However, backward compatibility is not total. Often when a drive type is first introduced, standards for data compression, etc. have not been established.

- **Age and Storage Conditions**. Electromagnetic and physical deterioration and even gravity can cause mis-reads. Manufacturers specify temperature, humidity and other storage requirements that are often not met in the real world. In the best of cases, media that may be perfectly suitable for day-to-day backup is inadequate for archiving.

- **Recording Methods**. Even though the physical medium may be the same, various drive models can record at different densities and using different schemes for data compression. For example, let's examine the popular IBM 3480/3490 family of tape systems:

  The 3480 tape drive records data on a ½" wide cartridge. 18 tracks of data are written concurrently side by side ("horizontally"). Each track is written "longitudinally" at a density of 37,871 bits per inch. The capacity of a given tape also depends on the size of each block of data, since the gaps between blocks also take up physical space. The maximum capacity of the standard 3480 tape cartridge is just about 200 megabytes. The 3490 drive is identical, except that each block of data is automatically compressed before writing to tape, and uncompressed on reading. This provides a nominal 2 to1 increase in capacity. (The actual increase depends on the nature of the data.) The 3490E drive records 36 tracks at a time, but is backward-compatible for reading 18-track tapes. It also supports compression, so that its nominal capacity is 4 times that of the 3480. In spite of a low storage capacity by modern standards, this type of tape is a standard at the National Archives because of its wide popularity, good archival qualities and rugged construction.

- **Operating System/Filing System**. The *OS* of the originating system (Unix, DOS, Windows, DEC VMS, Wang VS/OS, MacOS, etc.) usually also implies filing system[3]--the way files are stored and identified—for example, how long its name can be; what kind of characters may be included in the name; how disk drives and directories are organized; whether and how creation date or date of last access are saved; etc. Whether these conventions conform to—or can be made to conform to—the standards of the

[2] A decade or two ago, diskettes were much more predominate than today for offline storage. For a discussion of the many disparate media and formats of that time, one may refer to the article *"Six Reasons Why This Disk Won't Work in That Computer"*--available from the author on request.

[3] FYI - In the 70's and early 80's, virtually every dedicated word processing system (and there were dozens of them) had its own filing system.

archiving organization is a key issue.

- **Backup, Exchange or Archiving Software**. The next layer to be addressed when retrieving data from a tape is the type of program that was used to write the file to the tape. Historically, these programs have been written for one of three different purposes: (a) operational backup, (b) exchange of data with other systems, (c) archival storage for medium or long term. Some very popular utilities, such as Unix TAR are used for all three purposes.

  Most proprietary operating systems such as DEC VAX, Wang VS, IBM AS/400, Windows NT, etc., contain a standard backup program. These generally produce a proprietary tape format that is optimized for fast convenient <u>backup</u> with no mind toward being conveniently <u>read</u> on other systems. Further complicating things is the fact that there are a variety of popular third-party backup programs (e.g.- Cheyenne ArcServe). Systems without the same backup software generally cannot read those tapes without sophisticated conversion programs.

  Perhaps the most pervasive misconception is that the media and software used for backup are also good choices for archiving. Law suits and FOIA requests often provide a belated wakeup call, as we'll see in some later examples.

  The tapes that are easiest to handle are those intended for <u>exchange</u>—examples include ANSI and IBM "standard labeled" tapes. These standards are intended to persist across disparate systems and from one generation of software to the next. However, these formats often do not include related "metadata", such as the file's original disk location.

- **Application File Structure**. Within any file system, each *application program* such as Lotus, Dbase, or WordPerfect, may have its own method of internal file layout. Note that a Multimate document may not necessarily flow sequentially within its DOS/Windows file, where WordStar 3.x documents, on the other hand, flow nicely along in order. To create a successful conversion or viewing program, it is essential to understand the structure superimposed by the application program as well as the filing system.

- **Application File Encoding**. Coding conventions involve the general methods of describing printable text characters as well as simple control functions. There are two main conventions. The first, ASCII, the American Standard Code for Information Interchange is used by most mini and micro computer packages. The second, EBCDIC, is the Extended Binary Coded Decimal Interchange Code, devised by IBM for mainframe computers. As far as we know, only one PC-based WP package used EBCDIC as a base: IBM's Displaywrite.

  However, these coding conventions only serve as a base. Individual systems, such as different word processing programs, build on them in different ways. Not only the codes themselves, but their positioning relative to the text will vary widely from one application to another. For example, the way to underline a word may differ from one WP system to another.

  "Data" (non-text) files can contain many different ways to represent characters, numbers and dates. A few of these are integer, packed decimal and floating point—each of which come in a variety of flavors.

<u>War Stories and a Few Victories</u>

<u>FEDERAL RECORDS—A RAY OF LIGHT.</u> At the Center for Electronic Records, *"APS"* systems have been implemented to convert, preserve and catalog permanent federal records including such items as Iran-Contra era White House backup tapes and Warren Commission findings. The development of APS resulted in the ***1996 GSA Technology Excellence Award*** for section leader Fynette Eaton.

<u>A HORROR STORY</u>. Another federal agency was migrating to a new system. They converted several thousand "legacy" documents at one installation, through an on-line transfer, and then proceeded to decommission the legacy system, without making a final backup tape. Unbeknownst to the operations staff, the conversion software license had expired and <u>only the first page</u> of every document had been converted. If a backup tape had been made (and possibly used as the conversion medium), the process could have been re-run at any time.

THE IRS AND COMPANY X. A multinational corporation had accumulated data on about 1,000 reels of 9-track computer tape. The format of the data was not very well documented, and some of the tapes had already developed bad spots. The IRS demanded that the firm preserve the data for possible future inspection. The firm's solution: (a) preserve the data on CD (two copies)—this provides for much easier access as well as better shelf life; (b) the data was preserved on CD in both it's original structure and as converted to ASCII; (c) a software tool was provided to re-convert any of the original data as needed, should it be discovered that the original documentation was incorrect. This pretty well covered all possible future contingencies.

Law Suits Make Files Suddenly Important

FOIA CASE—STATE DEPARTMENT RECORDS. This case required that the State Department, via the U.S. Attorney's office, disclose specified records to the public. (The object of the suit was to find out whether a certain presidential candidate had traveled to the Soviet Union in his student days.) The original programmer had long since left federal employment without documenting the file layout. It was deciphered and the data was put it in a form to be presented to the courts.

POLYMER PATENT SUIT. Recently, a petroleum company needed to access suddenly-vital historical data stored on about 300 reels of backup tape from a "legacy" computer system in a proprietary word-processing format. The hundreds of thousands of electronic documents had to be converted to a more current word-processing format. The information needed to be placed on a more convenient medium (CDROM) for review by the court and attorneys for both sides. By conventional conversion methods, the process would take MONTHS, and the court (the famous "Rocket Docket" jurisdiction) would have none of that. This was complicated by the second problem: many of the tapes had been sitting on a shelf for decades, and had physically and magnetically deteriorated.

WHITEWATER--"VACUUMED" FILES. A Little Rock law firm exercised "due diligence" by retaining specialists to examine old disks thought to contain files related to the Whitewater investigation. Sure enough, several deleted files were discovered, recovered and also converted to a more compatible format.

ANOTHER FOIA: TECHNO-EXPORTS TO CHINA. Another agency has recently been ordered by a federal judge to turn over computer records from 1991 through 1996. These are on three different kinds of physical tape media (200+ tapes), recorded in a variety of densities, using four different backup programs, containing email and word-processing files from three different generations of application programs. Undoubtedly the court will require rendering of the information into a common form for searches and further perusal.

Ensuring Future Access to "Old" Data

A promising development for the future is the new MS66 standard being developed by ANSI and AIIM. To meet this standard, vendors of backup or archiving software need to create metadata that completely define their tapes' filing system. This would ostensibly enable one to write an application to retrieve files from the vendor's tapes should the vendor's own software be unavailable or impractical to use in the future. This could bring some degree of order—and the prospect of a greater degree of certainty of future access.

However, this does not address the existing universe of legacy data[4] that is the focus of this article. What to do about that?

Preserve, Preserve, Preserve. Hippocrates told his medical students "First, do no harm." The experienced computer archivist has an analogous first principle: "First, preserve that data." (Presuming, of course, that one's historical electronic records are of value.)

[4] Except for the possibility that one might develop tools to convert legacy backup tapes to a compliant format.

If only one copy of important information exists, it behooves the custodian of that material to make a second copy *by duplicating the first*. This is important, not only in the value of the second copy, but also by virtue of confirming whether or not the first copy was readable. To be completely certain of the validity of the second copy, one may elect to take the trouble to compare the two. This process is standard at NARA's Center for Electronic Records. (Some folks make a set of backup tapes, and then do another backup to create a second set for off-site storage. The logical flaw in this is that it confirms the readability and validity of neither set.)

This is not to imply that the second copy need be on the same medium. One can elect to copy an "image" of a legacy 9-track reel to CD, for example, for reasons of shelf life and storage space. However, one should retain the ability to recreate the original tape(s) as needed.

Convert, But Wisely. Assume you have important records in an older format or on older media than that which you currently support. Let's use Wang word-processing files as an example. There are three choices: (a) keep a "legacy" system up and running that can access that old data and convert/transfer documents to your new computer as needed; (b) bulk-convert all legacy files to a more modern standard (e.g.- Microsoft RTF); (c) acquire a conversion tool that runs on the new system, but can read/convert your legacy tapes in the future.

The choice is often not a simple one, combining budgetary and technical realities. For instance, choice (b) might be the best bet—but only if it is economically justified, and if you can be completely confident in the accuracy of the conversion. (Remember the "horror story" from above.)

For records that are both important and complex, one should consider preserving both the original and the converted version of each file. Sometimes a subtle error in conversion may not be discovered until much later. With the original intact, a remedy is always possible.

Transfer and Preservation Media. If one physically delivers tapes to a typical outside storage facility, the medium both for transfer to another "custodian" and for preservation is identical. If the new custodian copies to another kind of tape (ala NARA) then of course the two media are not the same.

Increasingly, the preferred medium for data transfer is the Internet. Of course, business and government agencies transfer lots of data among and between themselves. There are also companies now offering offsite incremental backups via the 'Net. Privacy and assurance of performance are key issues.

For the large volumes of legacy files sitting on government and industry shelves, the 'Net might be the way to transfer them to their ultimate custodian—if these criteria can be met:

- Files can be transferred or converted quickly and reliably enough from the legacy medium to a computer with Internet access.
- One can be sure that the file arrived at the new custodian (and when); signed return-receipt.
- Further assurance that the file has not been modified.
- Confidence on the part of the receiver that the sender is authentic.
- Confidence on the part of the sender that the file will only be read by intended recipient.
- All transactions must be logged; with easy reporting.
- The mechanism should be optimized for larger file sizes than regular email.
- The process must be easy to operate and administer.
- It must be cost-effective (of course). There's the rub.

The component technologies have been available for some time now—encryption, authentication, file transfer, virtual private networks, etc. I believe things are falling into place; we're approaching the point where all of the criteria will be met, so that the Internet, along with media preservation and conversion can help us defeat the "Phantom Menace".

Further information on these topics is available via the Muller Media Conversions web site at http://www.mullermedia.com.