



This article appears in the December 2004 issue of PRISM's *In Focus* magazine.

## *You're ready for that data. — Is it ready for you?*

By Chris Muller  
[www.mullermedia.com](http://www.mullermedia.com)  
Copyright 2004

Everybody's talking about compliance these days. But there are *all kinds* of reasons for needing to get at your company's accumulated data, including litigation, disaster recovery, research, etc. We have neither the space nor the mandate to go into them all here. (Thank goodness!)

There are three common threats to "Data Readiness".

- Loss of data due to drive crashes, operator error, damaged tape and disk media, viruses and the like.
- Failure to back up current data, (or to test restoration procedures). This is particularly true of small and medium-sized companies where the equipment, staff and procedures for rigorous backup methods can be overwhelming.
- Inability to access older data due to incompatibility or media deterioration. The first two items have received at least *some* attention in the press, but this category is often ignored until it's too late. So we'll focus on this issue--something I like to call "*the Phantom Menace*".

**TIME** and its minions are a menace to your data.

**TIME** and aging storage media cause "decay" of valuable records on the shelf.

**TIME** and the inevitable migration to new computers, software and media render older files incompatible and unusable on the new systems.

**TIME**, downsizing and outsourcing can lead to undocumented programs and data that are difficult to decipher.

*Note: A greater part of our information resources are threatened by **TIME** than by terrorists and hackers.*

### **The Onion: Obstacles to Long Term Information Access**

Let's review the underlying considerations that confront those trying to deal with long-term data availability. It's a bit like peeling back the layers of an onion:

- **Drive Type.** The great bulk of accumulated data is on tape. Some drive types have become obsolete (e.g. - nine and seven-track reel-to-reel). Others have seen a fairly orderly progression 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, certain standards have not been firmly established, leading to more incompatibilities.
- **Media Age/Storage Condition.** Electromagnetic and chemical 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. All these factors worsen over time. Media that may be perfectly suitable for day-to-day backup may be inadequate for archiving.
- **Recording Method.** Even though the *physical* medium may be essentially the same, various drive models can record at various densities and using different schemes for data compression. Recording density has a horizontal component (number of tracks) and longitudinal ("bits per inch" times the length of the tape). The capacity of a given tape also depends on the size of each block of data, since the gaps between blocks take up space.
- **Operating System/Filing System.** The *OS* of the originating system (Unix, MVS, DOS, Windows, DEC VMS, Wang VS, etc.) usually also implies one or more *filing systems*--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. The older conventions generally do not conform to the standards of modern systems.

- **Backup, Exchange or Archiving Software.** The next layer to be addressed is the type of program that was used to write the data 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 popular utilities, such as Unix TAR are used for all three purposes. Most operating systems contain a standard backup program. These generally produce a proprietary tape format that is optimized for fast convenient backup with no mind toward being read on other types of systems. Further complicating things, there are a variety of third-party backup programs. Some of these include converters for popular competing products. In many cases, though, expensive conversion tools are needed if you no longer have the originating system.

*Another misconception is that the software used for backup is also adequate for archiving.* Older tapes that were intended for archiving or exchange or are often easier to access than “backup” tapes. Examples include ANSI and IBM “standard labeled” tapes and Unix TAR format. These standards are intended to persist across disparate systems and from one generation of software to the next. However, these formats are often less efficient than proprietary packages, and do not include complete “metadata”, such as a file’s original owner and disk location.

- **Application File Structure.** Within any file system, each *application program* may have its own method of internal file organization. The native formats of sophisticated applications generally do not flow sequentially within a physical file. 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. To ensure future access to archival data, one should consider “exporting” such files to simpler formats, and/or producing human-readable report files based on the data.
- **Application File Encoding.** Coding conventions are methods of using numbers to represent printable text characters, formatting functions, and yet other types of numbers. There are two main conventions for text. ASCII (American Standard Code for Information Interchange) is used by most minicomputer and PC packages. EBCDIC (Extended Binary Coded Decimal Interchange Code) was devised by IBM for mainframe computers. However, these coding conventions only serve as a base. Word processing programs, for instance, build on them in different ways. Control codes and their positioning relative to the text vary widely from one application 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 comes in a variety of flavors.

### Ensuring Future Access to Older Data

It’s a tough call. What to do about that lode of legacy data that’s just sitting in a vault? Are we paying storage fees for material that could turn out to be useless to us? Does it have sufficient value that it’s worth being proactive, or should we wait in hope that we’re not forced later to take sudden and expensive action?

**Preserve, Preserve, Preserve.** Hippocrates famously said, “First, do no harm.” The analogous IT principle: “First, preserve that data on ubiquitous and long-lived media.” *Unfortunately*, the equipment and know-how to deal with older data is becoming increasingly rare. *Fortunately*, many older computer tapes have low capacity relative to CD and DVD, which are inexpensive as well as long-lived. One can preserve “images” of such tapes to Disc, improving both shelf life and physical space usage. Of course one must retain the ability to access the data from Disc when possible, or recreate the original tapes when needed. Since retrieving information at some future date from physically degraded or incompatible media could be very expensive, this can be the most cost-effective preventative measure. [This methodology is used for thousands of tapes at government agencies and large companies. Details provided on request.]

**Convert, But Wisely.** Assume you have important records in an older format or on older media than that which you currently support. Choices:

- (a) Keep a “legacy” system up and running that can access that old data and convert/transfer files to your new computer as needed. This option could be best in some instances, but can be expensive in floor space, maintenance fees and staffing.
- (b) Keep one’s fingers crossed and ignore the issue. Cheapest to start with, but certainly riskiest and potentially disastrous. Sadly, this is a frequent choice.

- (c) Bulk-convert all legacy files to a more modern standard. Perhaps best, depending on the price. Odds are, 90% of those old files will never be used—but which ones? Even so, don't be in a hurry to throw away the original data; horror stories abound<sup>1</sup>.
- (d) Preserve the raw material on long-lived ubiquitous media as described in the preceding section. Generally less expensive, and keeps open the "(c)" and "(e)" options for the future.
- (e) Do selective file conversions as needed using a conversion tool on the new system that can read/convert your legacy data in the future. If available at reasonable cost, this can be the best of all, perhaps combined with "(d)" so old tape drives are not needed.

### Law Suits, Compliance, FOIA etc. Make Old Files Suddenly Important

Many of us enjoy hearing "There but for the grace of God go I" sort of stories. Here are just a few that I've encountered which illustrate some of our IT foibles. In most of these cases, lack of documentation made things much harder for everyone. *Always save pertinent documentation along the data to which it relates.*

FOIA Case—State Department Records. 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. ...

Whitewater--"Vacuumed" Files. A law firm exercised "due diligence" by retaining specialists to examine disks thought to contain files related to the investigation. The system that created the disks was long gone. Sure enough, several "vacuumed" files were unearthed. Hmm—are these first two stories somehow connected? Lest we seem too partisan here, see below.

Watergate/Nixon Records. In the early 70's the Whitehouse had what was for then a pretty sophisticated computer system. (Would compete poorly with today's hand-helds.) There was a one-of-a-kind database containing presidential appointment calendar, contacts and other notes. Researchers had been itching to get at it for years. The data was saved on tapes in a form that deviated from subsequent IBM standards, and by the year 2000, the tapes had physically deteriorated. The "*Phantom Menace*" and all of its minions had really gone to work on this baby...

Giant Telecom Bankruptcy. Plaintiffs received cartons containing hundreds of jumbled, unmarked backup tapes created over a 3-year period. Except for an external barcode, all were visually identical. No documentation was available. (Some thought deliberately.) Using different software and different drive configurations. Some tapes went together in a series—a pretty common thing. Others, however, had also to be read in parallel ("tape RAID") in swaths of 2, 4, or 8 tapes. Many sets had one or more tapes missing. This all had to be worked out the hard way...

Frankly, doing this kind of work can be fun—if you're getting paid for it. But if you're the one doing the paying, taking reasonable precautions now may save tremendous amounts of downstream cost.

About the Author

*Chris Muller founded Muller Media Conversions in 1978. MMC is a leader in rescuing data from incompatible or obsolete formats. Chris has written several articles on conversion and preservation issues. Law firms, private companies and government often call upon his firm to "make sense" of obscure or difficult data files. For most folks, "WWW" is the world-wide-web. For Chris it stands for just three of the many interesting situations he's worked on: Watergate, Whitewater and WorldCom. :-)*

---

<sup>1</sup> A government agency we know used a conversion package on a legacy system to translate thousands of word processing files to PC. They then decommissioned the old system. Later they discovered that a fluke in the software had converted only the first page of most of the documents. Fortunately, they had preserved the original backups and were able to use a tape-based conversion package. *Don't throw data away unless you're sure it's been properly converted or replicated.*