

SENATE COMMITTEE ON COMPUTING AND INFORMATION SYSTEMS
SENATE COMMITTEE ON RESEARCH

Best Practices for the Preservation of Digital Data

(Informational)

Overview:

Research and academic data are stored by faculty members in a variety of media formats. Traditionally, these formats included use of notebooks, paper files, photographs and slides. Increasingly, however, the use of such media formats is decreasing in favor of storing data digitally. This transition has raised concerns and questions about how best to retain and preserve digital data.

Many of the issues concerning data retention are not unique to the used of digital storage media. For instance, security of storage media (i.e., data protection, access control) and media longevity (i.e., data preservation) are common considerations. In addition to preservation issues, all data custodians are expected to implement prudent and responsible safeguards to protect the confidentiality, integrity and availability of their data.

This report outlines several best practices for the preservation of digital data. A single approach or practice will not serve the needs of everyone. Each faculty member should evaluate his data retention requirements and take appropriate action to insure that his requirements are met. This document highlights several best practices, which are discussed in more detail below. These practices are as follows:

- Always keep more than one electronic copy of your data
- Choose storage media wisely
- Watch for media obsolescence
- Watch for disk format obsolescence
- Watch for file format and software obsolescence
- RAID systems and drive partitioning provide data safeguard strategies
- Establish a media migration policy
- Evaluate the value of your data on a regular basis
- Evaluate all policies and legal requirements for determining a minimal data storage duration
- Regularly test all components of a data storage system

Faculty members needing assistance to plan a data preservation strategy should work with college or campus IT staff to develop and manage a solution plan. Faculty members requiring assistance should start by calling the ITS Help Desk at 814-863-2494. Faculty members of the PSU College of Medicine should call the Hershey Medical Center IT Help Desk at 717-531-6281.

This document is intended for use by students, faculty and staff members who are managing data they generate for their personal use. This report should serve as an introductory user guide that highlights the main issues and the importance of data management. It is not intended to serve as a guide to professional individuals and groups that are responsible for management of Penn State institutional data, such as department or college Information Technology staff, members of Information Technology Services (ITS), Administrative Information Services (AIS) and University Libraries.

Best Practices:

1. Always keep more than one electronic copy of your data

- a. A “backup” digital copy is the most popular means of creating and preserving method to keep a second copy of your electronic data. The backup can be an identical copy or a copy in a compressed format. Most systems designed for data storage compress the stored files to conserve media storage space. The backup can be on similar or a different type of storage media.
- b. Store the original and backup copy of your data in separate locations. This practice is critical to lowering the potential of data loss due to security violations and environmental changes.
- c. When feasible, create paper files using high-quality paper kept in conditioned space.

2. Choose storage media wisely

- a. All types of media used for storing electronic data can and eventually will fail. Some types of media devices are more prone to failure than others. The tendency to failure for a particular type of storage media is expressed as the Mean Time Between Failures (MTBF). This time, usually expressed in hours, is the average time between failures for a particular device or system⁽¹¹⁾. The probability of failure increases with time and some media types have a longer MTBF than others.
- b. Relevant features in determining the appropriate data storage medium your purposes include data transfer rate, storage capacity, cost, network connectivity, and security. No data storage medium is ideal for all purposes. The media selection process is usually a decision based on a tradeoff between long-term stability (MTBF) of the medium, its cost and the data transfer rate. All factors should be considered when selecting a data storage media. Other factors may become important (size/space limitations, environmental factors) in some instances.
- c. Several experts have warned that CD’s are manufactured using varying processes, resulting in differences in quality and ultimately data storage longevity⁽¹⁻⁶⁾. Not only is storage longevity affected by the disk composition, storage conditions also affect the life span of a CD. Further, unlike original “pressed” CD’s, CD’s made with common CD “burners” should not be used as a long-term storage media.

Several reports have critically analyzed the factors affecting data retention on CD's and DVD's⁽⁷⁻⁸⁾. Experts recommend that CD's should not be used for data storage for periods greater than 2 years. It is ill-advised to keep primary or backup copies of data on CD's or DVD's.

- d. Rewriteable CD's (CD-RW) should never be used for backup or archival purposes^(2, 4). In general, they have less longevity than standard CD's (CD-R).
- e. Most experts agree that the best media for longer-term storage reliability is either magnetic tape or hard disk drives.

3. Watch for media obsolescence

- a. All types of storage media will eventually become obsolete. Even if the media remains intact, it may not be possible to use the media because some other components needed to read data from the media may become obsolete and/or unavailable. The obsolescence of drives and other reader devices can make recovery of data difficult or impossible. Floppy drives and older optical disk formats are examples.
- b. Users responsible for storing data should monitor that marketplace regularly for replacement hardware for media and other components (disk controllers, disks drives, etc.) in the event of a failure.

4. Watch for disk format obsolescence

- a. Tape and hardware data storage devices are formatted so that data is arranged in the medium in a manner that computer systems use to interpret data properly (FAT32, NTFS, etc). Data stored on devices formatted with older standards may become difficult to retrieve because of obsolete controllers and other hardware. If you have data on a device that was formatted using a near-obsolete or obsolete standard, those data should be moved to a device using a supported format standard.

5. Watch for file format and software obsolescence

- a. The file format is the particular way that information is encoded within the file so that the information contained in the file can be read by one or more application programs. The file formats that can be read by a particular application depend on the specifications of the individual software application. Data saved in some file formats can become obsolete as a result of the obsolescence of a software application and/or an operating system. As a practical example, a text file stored in an older version of WordPerfect (*.wpd) maybe more difficult to recover today than one stored as a Word document (*.doc) or as an ASCII file (*.txt). Consequently, saving data in the most universal format possible will help to avoid the problems associated with obsolescence of more limited formats.
- b. The file extension name (e.g., "doc") often indicates the file format type and the application used to write the file. It usually does not, however, indicate the version number of the software application used to write the original file. Such information may also be critical in recovering a file written in an older file format.

It is good practice to maintain a record of the file formats used (and version numbers) of files stored for archiving or backup purposes. These data files should be converted to newer formats before they become obsolete. This is also an important consideration when archiving or backing up data using a software utility that stores files or groups of files in a proprietary format. To save disk space, these utilities often compress the data. The original backup software, however, maybe required to recover the data. Data can be difficult to retrieve if the original backup utility is not available or the computer operating system is not available to run the utility.

6. RAID systems and drive partitioning provide data safeguard strategies

- a. A hard drive can be logically partitioned to create separate data storage locations on the same physical drive. This provides a convenient approach for creating a location for back-up copies of data files within the same computer. Although this approach is simple and it may prevent data loss due to a corrupt file, it does not prevent data loss due to hardware failure, theft or security breaches.
- b. A second backup approach is to use a RAID array. RAID refers to a Redundant Array of Inexpensive Disks⁽¹⁰⁾. RAID 1 is an array where files are duplicated (mirrored) on a separate disk. If a disk should fail, there is always a second copy of the files on another disk. A RAID controller in the computer often manages these operations automatically so files are written to two locations simultaneously when they are stored. In practice, the user and the operating system “see” only one disk and not several separate disks. As the cost of hard-disk space steadily decreases, RAID controllers are becoming a common option on more advanced personal computers to provide a simple approach to redundant storage. Although this method largely eliminates the loss of data due to drive failure, it does not prevent data loss as result of a corrupt file, computer theft or security breaches.

7. Establish a media migration policy

- a. Data storage experts recommend that users establish a media migration policy that not only takes into account the obsolescence issues discussed herein, but also the value of the data being stored^(1,9). Migration of data to new storage technologies should be undertaken when it becomes obvious to the user that
 - i. the new technology will become a dominant and industry-supported technology for the foreseeable future, and
 - ii. the costs of adopting the technology outweigh the costs of recovering data stored using the older technology.

8. Evaluate the value of your data on a regular basis

- a. The value of data is time-dependent. On an annual basis, data storage stewards should work with data owners to evaluate the value of owners’ stored data. At some point in time, it is likely that the sum cost of the hardware, personnel, and space required to store the data outweighs the value of the data being stored.

- b. The value of data can affect both how and where the data should be stored. There are many companies that specialize in providing on-line data storage solutions (e.g., Iron Mountain, www.ironmountain.com) when institutional storage solutions are either unavailable or insufficient to satisfy specific storage requirements. Users should critically examine cost requirements and security issues when determining where their data is stored.
 - c. Data can be stored in a manner that can make a data-value assessment an easier task. For instance, data files can be separated on storage media by both topic and by data of creation, as these are often key factors that determine their value.
- 9. Evaluate all policies and legal requirements for determining a minimal data storage duration**
- a. You may be required to store your data for a minimum duration of time by law or institutional policies. Many granting agencies, publishers, educational institutions and regulatory agencies mandate minimal data retention periods for data generated as a result of employment, contract or grant award. These guidelines and policies should be reviewed when developing a data storage and migration plan. Faculty members are encouraged to review data retention policies and guidelines with their College or campus Sponsored Research office.
- 10. Regularly test all components of a data storage system**
- a. Regular testing is the only way to insure that data can be retrieved from a storage system. If you store your data on an institutionally-managed system (central file server, etc.), you can be reasonably sure that the data is backed up and the system is tested routinely. Questions of this nature should be referred to your local IT administrator or the ITS help desk (814-863-2494). Regular testing is advised if you manage your own backup or storage system.
 - b. Testing of data recovery procedures should also include tests for data integrity. Check-sums, read-write verification tests and other procedures can be used to validate data integrity during data recovery and data migration.

Summary:

Many factors should be considered when planning a data backup and archiving solution. Much of the decision-making process starts with an analysis of the volume of data to be managed and the device containing the source data. For instance, a backup plan for data contained on a mobile laptop would be expected to differ from a plan where data are contained on a desktop computer located in a secure office space. The most probable cause for data loss is different in these two cases. Laptop users should always remain conscious of the high potential for data loss due to theft and hard drive failure. A desktop computer is less likely to be stolen and can be more easily configured with multiple disk drives for data duplication.

In all cases, the starting point is to devise a plan. For assistance, call your department or college IT administrator or the ITS Help Desk (814-863-2494). Faculty members of the PSU College of Medicine should call the Hershey Medical Center IT Help Desk at 717-531-6281.

References and Additional Reading:

1. Larry Medina, *CD's, Lies, and Magnetic Tape*, Computerworld, Jan. 10, 2006, <http://www.computerworld.com/blogs/node/1552>
2. John Blau, *Storage experts warns of short life span for burned CDs*, Computerworld, Jan 10, 2006, <http://www.computerworld.com/hardwaretopics/storage/story/0,10801,107607,00.html>
3. John Blau, *IBM expert warns of short life span for burned CDs*, PC World, Jan. 11, 2006, <http://www.pcworld.idg.com.au/index.php/index.php?id=1780671907>
4. Dahna McConnachie, *The burning issue is, CDs don't last*, PC World, Jan. 16, 2006, <http://www.pcworld.idg.com.au/index.php/index.php?id=1295482963;fp;2;fpid;1>
5. G.A. Marken, *CD and DVD Longevity: How Long Will They Last?*, Aug. 2004, <http://www.audioholics.com/techtips/specsformats/CDDVDlongevity.php>
6. Lee Gomes, *Beware the Fading Dye: Writeable CDs, DVDs Vary a Lot in Quality*, Wall Street Journal, June 21, 2004, (http://www.mama.com/technology/Tech%20News/WSJ_Portals2.pdf)
7. Oliver Slattery, Richard Lu, Jian Zheng, Fred Byers, and Xiao Tang, *Stability Comparisons of Recordable Optical Discs – A study of Error Rates in Harsh Conditions*, J. Res. Natl. Inst. Stand. Technol. 109: 517-524, 2004, (<http://www.itl.nist.gov/div895/gipwog/StabilityStudy.pdf>).
8. Fred R. Byers, *Care and Handling of CDs and DVDs – A guide for Librarians and Archivists*, NIST Special Publication 500-252, National Institute of Standard and Technology, Council on Library and Information Resources, October 2003, (<http://www.itl.nist.gov/div895/carefordisc/CDandDVDCareandHandlingGuide.pdf>).
9. Cornell University Library, *Preserving Cornell's Digital Image Collections: Implementing an Archival Strategy*, <http://www.library.cornell.edu/imls>
10. See: http://en.wikipedia.org/wiki/Redundant_array_of_independent_disks.
11. See: http://en.wikipedia.org/wiki/Mean_time_between_failure.

SENATE COMMITTEE ON COMPUTING AND INFORMATION SYSTEMS

Yaw Agawu-Kakraba
John P. Boehmer
John T. Harwood
Jeffrey C. Kuhns
Timothy M. Lawlor
Christopher P. Long
Willie K. Ofosu
Eric G. Paterson, Vice-Chair
Joy M. Perrine
Russell C. Scaduto, Chair
Richard J. Simons, Jr.
Eric E. Wertz

SENATE COMMITTEE ON RESEARCH

Bernard J. Badiali
Henry J. Donahue
James L. Frazier
Frederick E. Gildow
Cheryl J. Glenn
M. Kathleen Heid
Julia C. Hewitt
Ronald J. Huss
Ernest W. Johnson
James F. Kasting
Mark Kester, Chair
Amir Khalilollahi
Leah Y. Liu
Christopher J. Lynch
John M. Mason
Julian D. Maynard
Eric M. Mockensturm, Vice-Chair
Colin J. Neill
Eva J. Pell
Parag C. Pendharkar
Joseph C. Reese
David W. Richardson
Joseph M. Shostell
David B. Spencer
Jill M. Stahl
Joan S. Thomson
Marian Walters
Candice A. Yekel