

# **Media Choices for the Preservation of Digital Documents**

## ***Introduction***

On any given desk, stacks of diskettes and CDs compete for space with the piles of paper. Among paper records there is certainty about whether the information contained by the paper is intact because one immediately recognizes its integrity with a quick glance. In contrast, information on diskettes and CDs is not so readily ascertained because one needs to access the information through a secondary medium (a computer). This is the basis of concerns from the cultural heritage community about the preservation of digital documents. Although strategies and techniques for preserving traditional documents and artifacts are relatively well-known, the digital realm is largely unknown ground, leaving many uneasy about the prospects of vital digital artifacts surviving for more than a very short period of time. However, more and more documents are being created using digital tools, transmitted over digital channels and stored on digital media. These documents and artifacts may never be used as paper documents or as tangible works of arts made of traditional materials. Given this, it is important to deal with the issues of the durability of digital information.

## ***Current Research Strategies for Digital Preservation***

Current research has identified two key areas forming the foundation of digital preservation: the persistence of data on the physical media (hardware) and the ability to read the formats in which the data is stored (software). Both problems are interrelated but because independent decisions can be made about each, the result is a potential array of combinations that must be considered. This is compounded and complicated by the rapid obsolescence of both media types and data formats as the technology industry

moves forward with the creation of new media and formats for mass markets.

To solve these problems, three strategies have been identified: *migration*, *emulation* and *refreshing*. Migration and emulation address the issues of the software obsolescence either by changing the format into a newer, more modern format or by recreating the old viewing environment within the new environment. Research has concentrated on the risks involved with moving from one format to another and on building of infrastructure to support either migration and/or emulation.

Refreshing, on the other hand, focuses on the hardware and

...involves periodically moving a file from one physical storage medium to another to avoid the physical decay or the obsolescence of that medium. Because physical storage devices (even CD-ROMs) decay, and because technological changes make older storage devices (such as 8 inch floppy drives) inaccessible to new computers, some ongoing form of refreshing is likely to be necessary for many years to come. (Jackson, 2002).

Where the digital collection is small, *refreshing* is a relatively minimal activity with minimal resource expenditure. However, for large collections, this can be a significant time and resource investment. Thus, *refreshing* is best done within the medium's lifespan while maximizing the interval between each refreshing.

Unfortunately, determining the lifespan of current media is difficult because longevity claims can vary widely. For example, Jeff Rothenberg, in his report to the Council on Library and Information Resources, suggests that media lifetimes in general

may only last 5 years, while manufacturers, like Kodak make specific claims on their CD-R media, arguing that media may last 100-200 years. More cautious estimates, like those of John van Bogart at *Imation* suggest that digital media in general may last 10-30 years. Because simple calculations of lifespan vary so widely, these numbers cannot be used to readily determine the selection of a medium for digital preservation.

### ***A Strategy for Assessing Media Suitability for Archival Purpose***

The selection criteria for archival media should reduce the risk of loss compared to other media and facilitate the active management of the information.

To evaluate risk mitigation, one should consider:

- Is the medium physically durable? Having a hard casing, a protective door or other similar features will ensure that damage through poor handling will be minimized.
- Is the medium vulnerable to environmental factors? Media which can be stored in a greater range of environmental conditions will stand a greater chance of surviving in less than optimal conditions.
- Is the medium commonly available or is it highly specialized? By being commonly available, the medium should withstand technical obsolescence more easily than a specialized format.
- Is the medium write-once only? This will ensure that it won't accidentally be overwritten
- Overall, how many critical points of failure does the medium have? Media are more vulnerable when they have a greater number of "critical points"(technological obsolescence, physical vulnerabilities, mechanical failure points).

To evaluate the potential for active management one should consider:

- Does the medium require a sophisticated technological infrastructure? Media that are simpler to use and to deploy mean that a broader selection of staff can be utilized to maintain the collection.
- What is the capacity of the medium? Media with greater capacities reduce the amount of time necessary to perform refreshing activities because they require less media swaps to complete the task.
- What is the cost of the media? Of devices required to read the media? Active management should involve multiple copies to ensure redundant backups in case of media failure. Cost can be a factor in the ability to provide multiple copies.

### ***Media Formats and Technologies***

Currently, there are two primary technologies used for digital storage: magnetic and optic. Magnetic media come in a number of formats, including the floppy disk, tape systems, floppy or removable disks and hard drives. They all rely on magnetic particles in the recording substrate that change direction in the presence of a magnetic field. Optical technology including CD-ROM, CD-R/W, DVD-ROM, DVD-R/W and DVD+R/W are read using a laser beam which reflects the light from the surface of the disc in areas of differential reflectivity. Reflectivity results from a physically pitted surface on manufactured optical mediums like CD-ROM and DVD-ROM or from a dye-coated substrate that is “written” on exposure to a higher intensity laser beam.

A third technology, flash memory cell technology, has been gaining popularity over the last few years, primarily in devices such as digital cameras and PDAs. Flash memory technology uses cells which switch on/off electrical voltage without requiring power for maintenance, in a similar manner to computer memory.

Currently there are a wide variety of choices for flash memory including compact flash cards, secure digital cards and USB flash drives. But, this media has not yet been evaluated for its longevity and use as an archival medium.

Traditionally, magnetic tape systems have been a common choice for archival media in data centers, and their durability has been well studied. Tape systems are usually deployed within a networked environment and require considerable technical support. Schedules for the retention of information are usually set up according to policies generated by technology staff rather than archival principles. Typically, only large institutions can afford such systems, limiting their usefulness in the cultural heritage community.

Removable magnetic media in the form of floppy disks and larger format magnetic media like the Zip disk are not commonly used as a storage format any longer. In fact, current discussions about removable magnetic media tend to focus on recovering old data stored on magnetic media because of technological obsolescence. As an example of rapid obsolescence, the Zip disk was introduced ten years ago and rapidly gained favor in a number of fields, especially the design and publishing industries. However, as writeable CD media gained traction in the marketplace, the advantages of the Zip disk quickly disappeared and what was once commonplace, now is virtually impossible to find.

Hard drive media are self-contained units, usually internal to a computer system. Hard drives contain one or more magnetic platters which hold the data and a number of read/write heads. Although hard drives vary by capacity, interface and form, advances in hard drive technology so far have primarily increased hard drive capacities while reducing the price, with insignificant technological change. New developments in external interfaces like USB 2.0 and IEEE 1394 (Firewire) allow fast transfer to

external hard drive systems, allowing users to treat hard drives less as a fixed component of the computer system and more like removable media. Even desktop RAID systems (redundant arrays of independent drives) are now an affordable reality extending both capacity and the ability to recover from individual media failure.

There are several types of optical media with different formats which utilize the same technology. On read-only media (CD-ROM, DVD-ROM), the pits are physically molded onto the polycarbonate surface. On writeable media (CD-R/W, DVD-R/W, DVD+R/W), a layer of dye mimics the reflectivity changes that occur in the pits. Unlike the CD-R/RW formats, the DVD family of formats includes two competing standards, the “+” standard and the “-” standard. While newer DVD drives can read and write to both, older systems and systems like laptops may only allow you to read either one or the other. This format competition is reason enough to withhold judgment on the DVD family for archival purposes. Complicating the issue further, the next generation DVD formats include such rival options as the Blu-Ray and the HD-DVD. Therefore current discussions are mainly focused on CD-R/RW formats where the distinction between the “-R” technology and the “-RW” technology concerns capability – the “-R” standard is a write-once only format while the “-RW” standard allows for multiple writes and rewrites.

### ***Assessing the Options***

Individuals and small organizations primarily have available optical media such as CD and DVD technology or magnetic media such as hard drives. The capabilities of larger organizations may allow for a networked approach coupled with automated systems. These types of systems have been the focus of research at the National Digital Information Infrastructure Preservation Program

in the US but are beyond the scope of this discussion because of resource and technological expertise.

Currently, the primary choice for an optical medium is CD-R format. The write-once nature of CD-R's prevents accidental overwriting and their capacity is generally sufficient for large quantities of textual information. There are several different types of dyes in use for manufacturing CD-Rs, and recent discussions have focused on the best choices for dyes and substrates. A recent study (<http://nvl.nist.gov/pub/nistpubs/jres/109/5/j95sla.pdf>) concerning various dyes indicates that there may be some justification for using a more expensive (gold/phthalocyanine dye) CD-R, but the study's limited sample set suggests that the results may be premature. Since optical media such as CD-Rs are separate from the reader/writer devices, multiple copies can be generated at a lower cost.

This separation between media and reader also reduces the critical points of failure. Mechanical failure of the reader/writer devices does not impact the individual media so loss due to mechanical failure is largely minimized. The reading/writing mechanisms use light and no physical contact is made with the surface of the media. Although CD-Rs are vulnerable to light exposure, they are not affected by the magnetic fields generated by a computer/reader. Importantly, widespread use of this media and reader/writers strongly suggests that this technology will be supported in the future.

Risks in choosing CD-R media include the inability to change information once written, potentially leading to confusion between various versions. Capacities are also limited for some file sizes such as high resolution digital images and video, potentially resulting in a large number of disks. CD-R disks are also vulnerable to damage during physical handling.

The hard drive itself is an option for archival storage because virtually all digital documents are created on a hard drive and the equivalence of size simplifies the storage process. One can simply pull the old drive out and put in a new drive to start afresh.

Their self contained and hard casing will protect against environmental fluctuations within given operational parameters. They also cannot be accidentally scratched or otherwise damaged because they are incased in a computer case or an external enclosure. Given the longevity of the basic technology and multiple manufacturers, it is reasonable to assume that the media will be supported for some time to come, also reducing risk.

However, hard drives are susceptible to mechanical failure and this can cause data loss because the media is integrated into the device. The rotating spindle can seize or be displaced, or the platters themselves can be damaged by the contact between the read/write head and the platters or collision of the platters due to shock. A power surge can damage the drive, potentially burning out the electronics built into the drive, and thus destroying the data itself. External magnetic fields can also interfere with the drive and cause data loss. Also, hard drives, by design, are meant for read/write operations and accidental overwriting can be a common occurrence.

However, hard drives excel in their ability to promote active management. Data access is quick and allows activities like refreshing to be performed faster. In addition, size capability allows for storage of large collections on a small number of hard drives. This utilizes a small number of operations in handling a large quantity of information, freeing the user to perform other tasks while the system carries out the refreshing task. Additionally, larger files such as video or sets of high resolution images may require the memory of a hard drive for complete record storage.

## CONCLUSION

Often, the best advice is to choose more than one medium for storage and create multiple copies, ensuring that at least one copy will be readable. One approach is to use dedicated external hard drives for primary storage and to segment new data into CD-R sized blocks that are backed up onto CD-Rs as each block is filled. Finally, once a hard drive is full, it is pulled from service and acts as a backup copy. The CD-Rs are then used to access the data in a read-only fashion and when changes are made, the new versions are placed onto a new hard drive and backed up again onto CD-R. Although the issue of multiple versions can cause confusion, there are software solutions that assist in the management of multiple media and multiple versions. From this view, digital media preservation becomes less an issue of choosing a specific medium and more centered on an integrated strategy which takes advantage of the benefits of each component.

Tim Au Yeung  
Manager, Digital Object Repository Technologies  
University of Calgary

Contact: [ytau@ucalgary.ca](mailto:ytau@ucalgary.ca)