Storage media

This is a sample chapter from the forthcoming book,* Files that Last, *by Gary McGath. Copyright 2012 by Gary McGath, all rights reserved.

Say see you later to your data
And sing an ode to dead code.
Any drive can make you hate her
When a backup is owed.
Bill Roper, "Hard Drive Calypso"

Files are no more durable than the media they are stored on. What's the best alternative if you're concerned with longevity: tape, hard drive, flash, DVD, or something else? There isn't a clear answer, since each one has its own strengths and weaknesses, but some are better than others, and you can get good material or cheap junk in any medium. Whatever you choose, care and storage can make a big difference to its lifespan.

Sometimes you're on the receiving end of legacy files, and you have to deal with the medium they're on. In addition to current and upcoming media, this chapter looks at some of the older formats and their issues.

Compact discs

Compact discs (CDs) and their high-density relatives, digital video discs (DVDs) and Blu-Ray, are often a good choice for archiving. (These are always "discs," not "disks.") They're optical storage media, written and read by lasers. They're strictly passive objects, with no moving parts to fail, and they aren't susceptible to magnetic fields, so they can last quite a while. They've been common for a long time now, so it should be fairly easy to get drives for them for the next couple of decades.

CDs are good for small amounts of data; you can get about 800 megabytes on one CD. DVDs can store as much as 17 gigabytes (dual sided, dual layer). Blu-Ray can store as much as 100 gigabytes.

There are several different CD technologies. The original is the audio compact disc, or more formally CD-Audio. Commercial CD-Audio discs (other than small runs by some independents) are mass produced and can't be written with computer drives. A laser device creates a master, then a duplicator copies it, stamping pits in the recording layer. Discs made this way are very durable. Audio discs written from a computer are CD-R discs, and they're written with a laser that modifies points on a dye layer to change their reflectivity rather than burning pits. Dyes are subject to change over time, so even the best CD-Rs aren't as durable as true CD-Audio discs.

CD-ROM is similar to CD-Audio, but allows the storage of data instead of music tracks. It's formalized by the ISO/IEC 10149 standard. There are two data modes. The generally used one is Mode 1, which includes error correction. Mode 2 allows more data, but there's no error correction, so it's more subject to data loss. CD-ROM and other data CDs normally are formatted according to the High Sierra standard, standardized in ISO 9660 and ECMA-119. There are three levels of High Sierra. Level 1 is just a foothill, MS-DOS compatible, and rarely found. Level 2 and Level 3 both allow long file names which are case-insensitive; the maximum length depends on the operating system, but is generally at least 180 characters. Only letters, digits, underscores, and periods are allowed in the name. The main difference between these two levels is that Level 3 allows fragmented files.

Some extensions to ISO 9660 allow other characters in the filenames. A disc that takes advantage of this might cause problems on operating systems that don't use the same extensions. The Joliet extension is widely used and allows secondary filenames that can have Unicode characters. Non-Joliet operating system software can read the discs but will see only the ASCII filename. This could be a source of confusion, especially if the file name is in a non-Roman script.

The Universal Disk Format (UDF), which is compatible with the ISO 13346 and ECMA-167 standards. UDF supports Unicode file names directly, allows hard and symbolic links, supports a richer set of attributes, and is better suited for adding files to an existing disc. UDF is the usual format for DVDs and has probably surpassed ISO 9660 for CDs by now. Both formats should be viable for a long time, but UDF will undoubtedly win the standards battle until something else comes along.

How durable are CDs? Recorded commercial discs are the most durable because they're made by putting holes in the metal reflective layer. This is a modern version of the ancient and very durable information technology, engraving. For writable and rewritable discs, which are the only options computers normally provide, the storage method is more complicated and less durable.

A recordable CD is a sandwich of multiple layers. On top is the label layer, which just holds printed information. Next is a lacquer layer which seals off the vital reflective layer from air. The reflective layer is most often made of aluminum; on the best quality discs it's made of silver or gold, which are more resistant to oxidation. Under this comes the dye layer, which holds the recorded information. A disc is recorded by burning holes in the dye, which correspond to pits in the metal in a factory-recorded disc. Finally comes a thick transparent layer, usually polycarbonate, through which the laser can shine to read and write the disc.

Recordable discs use a metal reflective layer plus a recording layer which contains a dye. These are protected by a clear layer on each side. Usually the reflective layer is made of aluminum, which will oxidize if the seal of the protective layers is imperfect or breaks down. The dye can discolor over time, especially if it's exposed to light for long periods. The highest-quality discs use gold instead of aluminum, so they won't oxidize, but they cost a lot more. In the case of a dual-layer recordable DVD, there are two dye layers, with the first one being semi-transparent; the laser can be focused on the appropriate layer.

There are three types of dyes in common use. Cyanine or "gold-green" is the least stable. Pthalocyanine ("gold") and azo ("silver-blue") are significantly more stable. (Don't confuse "gold" dye with a gold metallic layer! You can't tell the metal of a CD from its color.) Some sources say that phthalocyanine is more susceptible to power variation than azo.

Properly stored, CDs can last a good long time. It's hard to put a figure on it, since the likelihood of failure depends so heavily on storage and handling conditions and the original quality of the disc. The most durable CDs, called "archival grade," are made with phthalocyanine and a gold reflective layer. The quality of the sealing is important; tiny defects can let air through and damage some bits. As long as the metal isn't exposed to oxidation, even aluminum can last quite a while. Manufacturers' claims for the best discs sometimes go as high as 300 years, which would let you drop a disc into a time capsule for your descendants, but 25 years is a safer estimate. If you get more than that, it's a bonus.

Surprisingly, a CD is less well-protected on its label side than on its playing side. There is only a thin layer between the reflective layer and the surface of the label. Scratches on the label may do irreparable damage, while scratches on the playing side can sometimes be smoothed out.

CDs should be stored at 10-20 degrees C (50-70 degrees F). Cardboard sleeves are OK if they're good

quality, acid-free paper. Jewel cases are better; they should have an internal tray or hub that isn't broken. Some sources suggest that the paper inserts should be stored separately from the discs to avoid acid contamination. Store CDs upright, not leaning or stacked.

Adding a paper label isn't a good idea for archival CDs. Printed labels have to be applied carefully. Leave wrinkles or apply the label off-center, and you could unbalance the disc and exacerbate whatever other problems it has. Even properly applied, they could cause acid damage or come loose after a period of years.

Writing is safer, if you use the right kind of marker. Writing on the label side with a ballpoint pen or other hard instrument can scratch it. The recording layer is closest to the label side, so a scratch from a pen could go through to the data. There's argument about whether regular permanent markers are better than soft-tipped markers made specifically for CDs. In principle, a marker with a strong solvent could get through the label and damage the recording layer. The "CD-safe" pens aren't that expensive, and they may in fact improve disc longevity. A good CD printer is a safe option, but they're expensive.

There are devices which can create bulk CD or DVD archives without human intervention — you might say a cloud of DVDs. Large operations might find these useful.

Avoid touching or scratching the disc surface. Don't use compressed-air blowers; these can chill the disc and cause damage. If a disc gets smudged or has data errors, wipe with a soft cloth radially (away from the center, not in a circular direction). If it's scratched, it may be possible to buff out a small scratch, but don't use cleaners with abrasives or solvents. There are scratch-removing machines which may be able to salvage scratched discs. If you can't read a disc even after cleaning, try it on a different machine, and be sure to save all the contents you can read.

This advice also applies to DVDs and Blu-Ray.

Digital video discs

The digital video disc (DVD) is similar to the CD in many ways, but its developers had the advantage of years of hindsight. The disc is also more encumbered with trade secrets and protection schemes. It was developed from the outset for file-structured data as well as audio-visual content. Even though optical disc technology was more mature when it got started, it has even more variants than CDs.

DVD-Video, the one you buy movies on, is the best-known version. The specification is expensive and requires signing a non-disclosure agreement. It isn't used for file storage, so there's no need to talk about it here.

The oldest data-oriented type is DVD-RAM, which was developed in 1996. It's hard to find drives which still support it. DVD-RAM discs are often sold in cartridges, enhancing their durability. As the name suggests, the disc allows random access for both reading and writing.

DVD-R, which came out in 1997, uses thinner disc layers, so that double-sided discs are possible. The name is pronounced "DVD minus R" or "DVD dash R." With single-sided discs, half the thickness is an inert layer that just brings it up to the requisite 1.2 mm thickness and protects from scratches. The capacity is 4.71 gigabytes. With double-sided discs, the layer structure is mirrored on both sides so that the disc can be flipped over. Dual-layer discs have a capacity of 8.5 gigabytes. The double-sided ones are rare for file data, since one disc that you can flip over doesn't really offer much of an advantage over two single-sided discs, and you can't write on it. The combination of double-sided and dual-layer is rare.

DVD-R is a write-once format; you have to record the disc all in one session. It uses a reflective dye, like the one used in writable CDs.

DVD+R ("DVD plus R") is an advance over DVD-R, because of better error correction. The physical structure is the same, and most computer drives today can read both DVD=R and DVD+R. Like DVD-R, it's a write-once technology. The capacity is virtually the same as DVD-R, and dual-layer and double-sided ones exist, and again the combination of double-sided and dual-layer is rare. The "plus" discs are a bit more expensive than the "minus" ones.

There are rewritable versions of both of these formats; not surprisingly, they're called DVD-RW and DVD+RW. They can typically be written about a thousand times, so they're for incremental storage, not for routine updating of files (and they'd be very slow for that). Unlike the write-once versions, they use a phase-change technology, and the phase change in the reflective layer is reversible, allowing rewriting, but compromising data durability compared to the write-once versions. Don't use these for archival storage.

Don't worry about region codes. They apply only to video DVDs.

How do CDs and DVDs compare in durability? There are a lot of factors and no clear answer. Let's compare best against best: high-quality CD-R against high-quality single-layer DVD+R. In favor of CD-R, the bits are less dense and the pits or dye markings are larger.The underside, which is the playing side, of a CD is protected by a thicker layer than a DVD. In favor of DVD+R, more sophisticated tracking and error correction are used, and label-side scratches are less likely to cause data loss. Which wins out in the end? We may know in twenty or fifty years. Choosing the best quality media and storing them under good conditions is the most important thing.

Blu-Ray discs

At this writing, Blu-Ray discs for data haven't caught on very quickly. DVD drives are still the norm, and some computers can read Blu-Ray discs but not write them. In spite of their greater density, they actually have some advantages over DVDs for data durability. Their prices have dropped significantly in the past few years and now cost about a dollar apiece in quantity.

The basic technology of Blu-Ray (abbreviated as BD) is similar to CD and DVD. A laser reflects off the disc, and pits or dye alterations change the reflectivity of a spot. Blu-Ray uses the full thickness of the disc, unlike DVD, which uses only half in a single-sided disc. (There are no double-sided Blu-Ray discs.) There is a two-nanometer hardcoat over the reading side, which provides extra protection against scratches and marks. The better quality DVDs also have a hardcoat, so this may not be a practical difference.

A single-layer disc can store 25 gigabytes and a dual-layer one stores 50. The newer BDXL format increases data density and allows more layers. The triple-layer BDXL disc holds 100 gigabytes, not because the third layer somehow has twice the capacity of the first two, but because the composition of the layers is different. Older Blu-Ray drives won't read or write BDXL.

Write-once discs are called BD R, and the rewritable version is BD RE (not RW). RE discs may use organic dye, inorganic alloy, or phase-change technology. The ones that use dye are considered the most durable, and again a gold reflective layer is best. As with the other disc formats, the write-once BD R offers better longevity than RE.

HD DVD discs

For a few years, HD DVD competed with Blu-Ray as the next-generation DVD format. It gradually lost ground, and when Warner announced in January 2008 that it was going exclusively with Blu-Ray, that was the end of HD DVD. Toshiba announced an HD DVD-RW drive, but it's not clear if any were ever shipped. In the unlikely event that you have data on HD DVD which you need to keep, get it off there and onto something else.

Summary of risks for optical discs:

- Scratches and warping can damage them.
- Oxidation can make data unreadable.
- Dyes can fade, particularly under light.
- Rewritable discs are less durable.
- Paper labels can unbalance discs or leak acid over time.
- Some CD file systems have restrictions on file names.

Summary of approaches to retention:

- Use discs with good construction and dye. A silver metallic layer is better than aluminum, and gold is better than silver.
- Use write-once discs.
- Mark discs with a CD marker.
- Stack them vertically in jewel cases in a dry place near room temperature, out of direct light.

Magnetic disks

Magnetic disk drives are the mainstay of computer storage. They're reliable, fast, and high in capacity. They're often recommended for long-term storage. However, they won't last forever, and they have their own special vulnerabilities.

A magnetic drive is made of precision components; the head rides a few micrometers above the disk, the disk is spinning at thousands of RPM, and if the two ever meet, the result is a "head crash," which sounds like a buzzsaw attacking the disk, and you can "say see you later to your data." If you leave a drive alone for a few years without plugging it in, components may get sticky, dust may settle in, and when you power it up again you may get a horrible noise and a whiff of smoke. A single dust particle can be bigger than the distance between the head and the drive, so it's like a boulder on the highway. The spindle might not spin properly after years of disuse.

Magnetic fields are a risk. It takes a hefty dose of magnetism to zap a modern drive that's safe in its box, but moderate fields can shorten its life over a period of years. Electrostatic discharges can damage both the the electronics of a drive and the data on the disk.

A disk drive is just a whirring box unless it has an interface. Once upon a time, SCSI (Small Computer System Interface, pronounced "scuzzy") interfaces were standard — at least if you can call an interface where everyone used a different plug "standard." Multiple devices could be daisy-chained, but the behavior was often finicky. There was a saying among hardware engineers: "SCSI is not black magic. There are fundamental technical reasons why it is necessary to sacrifice a goat at midnight in order to

get a SCSI device working properly." SCSI isn't found on many modern computers, but it's still possible to get a SCSI card if you need to connect to an old drive.

With modern disk drives, there are usually two levels of interface to a drive in its own box. There's the interface to the drive and the one to the box. A drive built into a computer only has one interface. The most common drive interfaces today are parallel ATA, also known as IDE, and serial ATA.

Parallel ATA can use three different types of cable and two different types of connector. The 40-pin connector is usual with 5.25 inch and 3.5 inch drives. The 80-wire cable uses 40 grounded wires alternating with the signal wires to reduce crosstalk and allow faster data rates. Some drives, usually smaller ones, use a 50-pin connector with a 44-wire cable. Four of the extra pins are used for jumpers and two are unused.

Serial ATA or SATA, as the name suggests, sends bits serially over a single line rather than in parallel; in spite of this, it's generally faster than parallel ATA. It uses a 4-wire cable; the connector has 7 pins, including three grounds.

These interfaces are specific to drives, and a long 80-wire cable would be clumsy and expensive, so external disk drives today are generally connected by cables that are suitable for a broad range of devices. USB has kept its leadership, with FireWire and Thunderbolt competing for market share.. Inside a USB drive enclosure you'll generally find a Serial ATA or SATA drive. With older drives the connection to the box may be some hard-to-find interface, but the drive inside is likely to be ATA. It's relatively simple work, for someone who's familiar with computer hardware, to remove the drive and put it inside a computer, or into a box with a USB-to-ATA adapter. In the future SATA drives may be around when USB is obsolete, and it will be possible to transplant the drive to a box that uses the latest interface.

Since hard drives are often built into computers, one way to avoid doing either is to save the whole computer. If it still works when you're ready to recover the data, then you don't have to worry about software compatibility. The downside is that there are more points of failure; either the computer or the drive can fail. If you take a 20-year-old computer out of its vault and it doesn't work, good luck finding replacement parts! There's something to say for a compact computer as an archival box, as long as you don't go for the cheap end.

Summary of risks for magnetic disks:

- Physical impact can cause irreversible damage.
- Strong magnetic fields and electrostatic discharge can harm data.
- Physical interfaces may become obsolete.
- Disks can freeze up in long-term storage.

Summary of approaches to retention:

- Keep at room temperature or slightly cooler, away from hazardous conditions.
- Deliver power through a surge protector or UPS.
- Avoid moving them while spun up, except for ones (e.g., laptop drives) built for mobility.

Floppy disks

Floppy disks are a magnetic disks made from thin and pliable material. They're an obsolete and fragile medium, but they were so popular that you can still run into them. There were three major form factors, successively shrinking.

The first writable floppies, made in 1972, were eight inches in diameter and were permanently contained in a soft envelope. They had a data capacity of 175 kilobytes (that's right, just 175 *thousand* bytes). Later double-sided and double-density 8-inch disks were developed, with capacity of up to a megabyte. Part of the magnetic surface was exposed to air and fingerprints, and disks could fail at any time. Different operating systems used different data formats.

The 5¼-inch floppy appeared in 1977. It was similar in construction and appearance to the 8-inch one. The first ones held about 110 kilobytes, depending on formatting; later ones could hold as much as 720 kilobytes. They had the same vulnerabilities as the bigger disks.

The 3½-inch floppy, first seen in 1983, emerged from several competing successors to the 5¼-inch disk. It enclosed the soft disk in a hard shell which protected it from direct contact and kept out most dust. The drive slid away a shutter to let the head touch the disk. This eventually developed into the high-density format, which can hold as much as 1.44 megabytes.

3½-inch floppies are more standardized and durable than their ancestors, though they're hardly preservation-quality media. If you come upon data stored on these disks and think it's worth keeping, you can buy an inexpensive floppy drive and, if all goes well, get the files off them. Don't leave them lying around if you're at all worried about losing the data.

Summary of risks for floppy disks:

- They're extremely fragile.
- The magnetic surface of 8" and 5 ¼" disks is open to the environment.
- Formats of older floppies aren't standardized.

Summary of approaches to retention:

- Handle with extreme care.
- Copy data off them at the earliest opportunity.

Magnetic tape

Tape is still used as a bulk backup medium by large installations. Its can hold a lot of data compactly, and good tape lasts a long time when properly stored. Its disadvantages are that it's slow and vulnerable to electromagnetic fields and formats aren't well standardized. It's generally used as an extra layer of backup with disk storage. The lack of standardization means it's not a good medium for long-term storage; it could be very hard ten years from now to read a tape made today. The 1960 census results were stored in a tape format which only one machine in the world can now read.

Tape needs to be handled carefully. A jam in a tape drive can stretch or break the tape. It's exposed to air and needs to be kept in a clean environment. It can quickly melt if subjected to extreme heat. Tape seals or cans should be used to protect unmounted reels.

Even trying to give a representative list of magnetic tape formats would be hopeless. Individual manufacturers each have long lists of incompatible tape systems. If you want to keep files around for

many years, just don't use magnetic tape. If you have some old tapes you're trying to get data off, good luck.

Summary of risks for tape:

- They're vulnerable to magnetic fields and exposed to the air.
- They can stretch and break.
- There is no standard format.

Summary of approaches to retention:

- Store in a controlled environment, using tape seals or cans.
- Be alert to hardware obsolescence.

Flash storage

Flash storage devices are small, lightweight, and convenient. Are they reliable for long-term storage? That's not clear. These devices are EEPROM (electrically erasable programmable read-only memory) chips combined with an interface that lets them plug in and interact with operating systems as if they were disk drives. They have no moving parts to wear out. However, they do have a limited life cycle if they are written repeatedly; a maximum of 10,000 writes is typical before the drive fails. For long-term storage this limit isn't very important, but the devices can fail for other reasons as well.

There's serious price competition in consumer-grade USB flash drives, so they're made cheaply. (They're sometimes called "memory sticks," but the name properly belongs to a less common Sony device.) As with any other technology, some are made better than others. The connector may fail. The unit may be vulnerable to static discharges; one with a cap or case that isn't easily lost is worth paying extra for.

There isn't a lot of information on how well flash devices last when stored for long periods. You can maximize their chances by writing a device only once and keeping it stored at a moderate temperature away from moisture, magnetic fields, and electrical discharges. It won't succumb to mechanical stickiness the way a disk drive can. Still, flash drives have a poor reputation for long-term storage. EEPROM storage depends on storing an electrical charge, and charges tend to leak out of anything that isn't stored in vacuum. I can't find any good hard data, but they're reputed to have a maximum shelf life of ten years, poorer than CDs or DVDs.

Another consideration is that a USB flash drive is wedded to its connector. If USB drives go out of style in ten years, it may be as hard to find a computer that will read them as it is to find one that will read a SCSI drive today.

Flash memory sticks are usually formatted using FAT32, though they may have any drive format.

SD cards, the things that hold the pictures in a digital camera, are based on the same technology as flash memory sticks, but they're less durable. They're for local storage and moving files around, not for any kind of long-term storage.

Because they're cheap and portable, consumer-grade flash drives have a valuable place in short-term backup strategy. They can hold extra copies of critical files and be stored off-site. You can keep one in your car and carry another in your backpack, purse or briefcase. But they aren't a good choice for long-term file storage.

Summary of risks for consumer-grade flash devices:

- Most are cheaply made and not durable.
- Connectors are built-in and may become obsolescent.
- Bits “wear out” with repeated writing.

Summary of approaches to retention:

- Use for moving files around or as an extra copy, not for long-term storage.

Solid-state storage

There are high-capacity storage devices based on storage techniques such as Flash and DRAM (dynamic random access memory). DRAM has a basic problem: the word “dynamic” means that the memory must constantly be refreshed by built-in circuitry, or it fades almost instantly. Battery backup is an absolute necessity; often it’s built into the device. Put it in a closet for a long time, and the battery will run down. A DRAM-based storage unit can’t even be considered for long-term preservation.

With high-capacity flash, the technical issues are similar to the flash memory sticks already discussed, but they’re much better made. They’re generally better-made than the memory sticks, but they’ll “wear out” with long use. How well a good-quality flash drive will last in storage isn’t well established yet.

If you use one of these drives, don’t defragment it. Unlike disk drives, they have no seek time, so it doesn’t matter if files are contiguously stored or not, and the operation will decrease the drive’s life a little. Some operating systems do regular defragmentation by default. Turn it off for solid state drives. As solid-state devices become more common, operating systems will be better optimized for them.

Solid-state drives are becoming popular on laptops, where they save weight and size, and they’re standard in tablets and pocket devices. Considering the bouncing these devices get, storage devices with no moving parts probably have the advantage.

I haven’t found any solid (no pun intended) information on this, but the nature of the technology suggests that drives left in a closet will gradually lose charge. There don’t seem to be any figures on how long it will take before this becomes a problem.

Summary of risks for high-capacity solid state devices:

- Bits “wear out” with repeated writing.
- Longevity when powered down is uncertain.

Summary of approaches to retention:

- Disable defragmentation and other system features that do unnecessary writing.
- Avoid leaving powered down for years.

Concluding thoughts

No currently available data storage medium is “permanent” over more than a few decades. There may be better choices in the future; one exciting prospect, still in the research stages, is built from iron nanoparticles inside carbon nanotubes. They’re changed from 1 to 0 and back by physically sliding

along the tube. The developers of this medium are talking about durability in the (Carl Sagan voice) *billions of years*. That's long enough not just for future generations, but future species. But we don't have it yet. For now, files will have to be migrated to new media every generation or so.

Unless you're with a well-funded organization that has a long-term charter, you don't have much assurance about how people will handle your files in the future, but a good choice of media gives you a chance that they'll still be readable when someone gets around to checking them.

For really large amounts of storage, in the hundreds of gigabytes or more, a self-contained disk drive is the best way to go. For smaller amounts, high-quality CDs or DVDs (and Blu-Ray discs as they become feasible) are a better choice, because they don't risk drive failure or electromagnetic damage and aren't dependent on a particular interface. It's safe to assume that drives which can read them will be around for another thirty years or more.

Either way, spend a little extra for good-quality media. It's worth it if you care about having the files last that long.

Fitting media with the strategy levels:

Level 1 (personal data): Magnetic disk drives are reliable and inexpensive. CDs and DVDs are good for making inexpensive archives. USB flash drives can hold extra copies in places where even a CD would be inconveniently big, but don't count on them for long-term storage. You most likely have at least one solid-state drive in a phone, tablet, or laptop.

Level 2 (small businesses and organizations): Not much difference from Level 1. USB flash drives are good for passing files from one person to another when sending them over a network is less convenient.

Level 3 (larger businesses and organizations): These organizations are likely to add magnetic tape backup. If large numbers of CDs or DVDs are produced for archiving, a CD/DVD mass archiving device or a CD printer might be a reasonable option.

Level 4 (critical data collections): The Level 3 options apply. CDs or DVDs probably have the best shelf life of any medium. Any discs used should be of the best quality, with a gold metallic layer, and stored with care. USB flash devices aren't appropriate here.