How Documents Can Be Stored Electronically?

It is helpful to have a basic understanding of the technology terminology when working with electronic documents. Documents that must be retrieved immediately are often stored to a Direct Access Storage Device or (DASD); this is your local fixed disk storage device on your workstation or severs. However, since this is very costly storage and the most volatile to failure, important documents should be stored to archival media. Archival media has a longer storage life, is removable, and is generally operating system independent. One of the primary storage mediums for electronic data today is Optical Disks. Optical Disks are storage media from which data is read and to which data is written by lasers.

A discussion of how optical disks are created is beyond this paper, but the following are excellent for those wanting to know more.

http://www.jegsworks.com/Lessons/lesson6/lesson6-9.htm
http://teachmecomputers.tripod.com/lesson_4.htm

There are four basic types of optical disks.

- **WORM -** This stands for Write-Once, Read -Many. With a WORM disk drive, you can write data onto a WORM disk but only once. The disk may then be read many times. WORM is the earliest technology, but it is not widely used today as newer technologies are capable of much higher storing densities and faster speeds.

- **CD-ROM -** Compact Disk, Read Only Memory, or CD-ROMs come with data already encoded onto them. The data is permanent and can be read any number of times, but CD-ROMs cannot be modified. Today, CD-ROMs are most often distributed by software manufacturers or information resellers. Because DVD media is capable of storing much more information, DVD disks are quickly replacing CD-ROM in all areas except music distribution.

- **CD-R, CD-RW, CD+R, CD+RW -** Compact Disks are specifications for CD media. The dash and plus refer to the write and read technology as well as the hardware specifications of the drive. In addition, technology deploys streaming technology and is considered faster and more reliable. CD is being replaced today by DVD disk because of the greater storage capacity.

- **DVD-R, DVD-RW, DVD+R, DVD+RW** - What originally stood for "Digital Video Disc" is now simply a general term for large volume data storage medium. Like CD, the dash and plus refer to the write and read technology being deployed. Today, we are also seeing the emergence of HDDVD and Blue-Ray DVD (and high density implementation similar to HDDVD), although these are primarily used today for distribution of high-density movie viewing. This technology will quickly emerge as a viable long term data archive storage platform as well.

How long will an Optical Disk last?

A common question is "How long will and an optical disk last?" A factory-pressed disk is very different from recordable media you record using your computer. With pressed disk the data is literally molded or pressed into the media and will not disappear unless the disk is physically damaged. Recordable disks use a dye that changes color or reflectivity when heated. There are different dye types commonly used in recordable disk such as phthalocyanine, azo, and cyanine, in particular--and they do not all have the same life expectancy and stability.

The studies we have seen suggest that properly burned one-time media (-R media, but not -RW media; see below) has an expected life of decades to possibly even centuries.

There was a study by the National Institute of Standards and Technology on the relative stability of different media here: http://www.itl.nist.gov/iad/894.05/docs/StabilityStudy.pdf

You can see some comparisons in the NIST study of the different dye types. But this study did not attempt to extrapolate the data to a life expectancy, although it did provide data about the relative stability of the different dyes and reflection layers behind them. The Optical Storage Technology Association's report, http://www.osta.org/technology/cdqa13.htm, suggests that optical recordable media will last 50 to 200 years. This observation is backed by quite a number of studies that I have seen done both by the media makers and others. However, some storage experts suggest numbers more in line with a life of only 2 to 5 years. This is based on quality of the material, as well as the disk drive that created the disk. There is much discussion on this as both vary widely. As we noted above, pressed disks will last a very long time. The materials are consistent and the image is pressed into the media. Disks created by your computer will vary greatly based on the technology applied and the knowledge and skills of the person creating the disk. The truth, we imagine, is somewhere in between. If quality disk are used and the computer is maintained regularly so that the lasers are in alignment, etc. Then a safe period of time might be 5 – 50 years. The real threat is having a device that can actually access the data over that period of time. Just imagine the problems reading a disk over 200 years old. The second hurtle is having software that can read the stored format over the same period of time. If you found an old WordStar file, as an example, stored on a 15 year old floppy disk today – do you have a program that is easily accessible to read that disk, assuming the disk is still good?"

Above we mentioned that both hardware and software affect the quality and the life of an optical disk, here is some additional information that may be of value to you:

The quality of the disk burner

A borderline defective burner can "under expose" the media to the laser beam, producing a seemingly good recording (at the time of burning) that will "fade" over time (failing weeks, months, years or decades sooner than it should have had the laser beam intensity been correct)

Recording speed

Fast burns (52X) are probably less stable than slower burns (16x to 32x), but you can burn media too slowly also. A very good analogy is to compare this process to photographic film and exposure levels. The dyes on a given media have a certain range of acceptable "exposures" and outside of that range, you can either under or over expose the media to the laser beam. However, mechanical jitter and certain other variables (largely a function of the quality of the drive) generally will be unconditionally worse at faster speeds.

Handling and storage practices

On a CD media, the data "exists" in a dye layer on the label side of the media. This can be scratched from the back (from the label side), which will literally and directly destroy the data. The front side is clear plastic but can also be scratched. While front side damage may make the data less readable or completely unreadable, the data is still intact and undamaged on the label side, and the scratches on the front can normally be removed by polishing the plastic. On recordable DVDs, the data is on a layer "inside" the media, but the media is a laminate of several layers and can delaminate, destroying the data. Flexing – even VERY minor flexing – is particularly bad at causing such damage. Recordable DVDs tend to fail from the outside in, so *you can increase your success rate and decrease the incidence of failures by not recording such media beyond 80% to 90% of capacity*, leaving the outside edge, where the failure rate is greatest, blank.

Labeling

The glues in adhesive labels or the solvents in pen-type markers, both applied to the label side (the side containing the data) can SLOWLY penetrate the reflective backing and dye layers and destroy the data. Therefore, for archival media, the safest policy is to not label optical disk itself at all. If you do label it, with either a label or a pen, you are, at best, taking a chance with your data (hint: it is safe to write on the clear inside hub (where there is no data at all) with a suitable pen that won't rub off).

Erasable Media

Rewritable, or RW media, is FAR less stable than one-time "R" media and should absolutely not be used for any permanent recordings of any kind whatsoever. There is no question that RW media can and does "fade". We routinely see "RW" recordings that are unreadable after periods of months to a year or two when there is really no other explanation for the failure. We have not recordable failure of "R" media, but this may be a matter of the age of the technology and how long it has been in practical use.

Traditional Digital File Formats

Likewise, we store data in many different formats depending on the purpose for which the data is being retained. The following is a quick definition of the most popular file formats deployed today.

TIF - The *Tagged Image Format File* is images stored as single bits or dots. TIF files can be compressed without losing image quality. A page stored in uncompressed TIF at 300 dots per inch takes about one million bytes of disk space. A compressed TIF page will take about 65,000 bytes. These estimates may vary widely depending on whether the data being stored in the file is text, graphics, or a combination.

PDF - *Portable Document Format* was created by Adobe and Apple and can actually be several different kinds of files. A graphic PDF has no ASCII characters, and the image is stored as a TIF file inside the PDF file. Another PDF format is a graphic image with ASCII characters. This is created when an image is translated via Optical Character Recognition (OCR). PDF images can be viewed on almost all computers using Adobe Reader or other software. The file size for PDF may vary widely; we generally use 23,000 kilobytes per page as an average. Again, the actual file size will depend on the compression settings, and the data, text, or graphics, being stored. We recommend you create some sample PDF files that represent the type of files you will be storing and create benchmark figures for your average page size in characters.

JPG - *Joint Photographers enhanced Graphics* files are usually used by digital cameras to store compressed color graphics without much image degradation. This format is not recommended for the storage of most business documents unless they are pictures. JPG is typically used for photo or images, as opposed to RAW which some high-end digital cameras generate. JPG files are smaller than RAW files with some pixel extrapolation to reduce the file size. The size of the file will often be determined on the pixel setting of the camera at the time the image was created.

GIF - *Graphics Interchange Format* file is a common file format used to store images such as line drawings, photos, documents, and simple animations. The GIF file format is widely used on the Internet because it uses compression to reduce the file size containing the images. However, some image clarity may be lost between JPG and GIF.

ASCII – *American Standard Code for Information Interchange* was agreed upon by a joint government and military commission several decades ago. Each letter, number, or symbol is stored as a unique number in a base 16 or hexadecimal numbering system. The software program arranges them on the screen so you can read them. An ASCII file storing one page is about 2,000 bytes in size. Microsoft has a similar standard called *Comma Separated Value* (CSV) format for files which are human as well as machine readable.

BLOB - *Binary Large Object* files are used to store images in a proprietary format. A BLOB image file must have special software to be viewed or transferred. Depending on the software, a BLOB file may take slightly less space than a nonproprietary format like TIFF or JPG.

XML and XPS – Our discussion of file formats cannot be complete without a brief review of XML and Microsoft's XPS standards. *Extensible Markup Language*, XML, is a markup technique to tag data elements in a file. It is a common misconception that there is a single XML standard. There are, in fact, many sub-XML standards. Accountants should familiarize themselves with *Extensible Business Markup Language*, XBRL. If you are in another industry such as Real Estate, Architecture, or Law, you have your own unique set of XML markup standards and should be familiar with those as well.

XML is a markup language for documents containing structured information. Structured information contains both content (words, pictures, etc.) and some indication of what role that content plays. For example, content in a section heading has a different meaning than content in a footnote, which means something different than content in a figure caption or content in a database table, etc. Almost all documents have some structure. A markup language is a mechanism to identify structures in a document. The XML specification defines a standard way to add markup to documents. So, what is a document for XML discussion purposes? For our purposes, the word "document" refers not only to traditional documents, like this one, but also to the myriad of other XML "data formats" we mentioned above. These include vector graphics, e-commerce transactions, mathematical equations, object meta-data, server APIs, and a thousand other kinds of structured information. It may also refer to data transactions such as Accounts Receivable, Accounts Payable, General Ledger, and so forth. Many applications from accounting applications to Microsoft Office produce XML formatted files today which allow data to seamlessly flow between disparate applications. For more in depth knowledge, we suggest you visit www.xml.org, www.xml.com and, for accountants, www.xbrl.org.

While Microsoft's *XML Paper Specification* (XPS) is relatively new, we believe it is important that you be at least aware of it. XPS describes electronic paper in a way that can be read by hardware, read by software, and read by humans. With XPS, electronic documents can be printed, shared, and securely archived. Microsoft integrated XPS-based technologies into the 2007 Microsoft Office system and the Microsoft Windows Vista operating system. However, according to Microsoft, XPS itself is platform independent, openly published, and available royalty-free. While PDF is still the dominate file format for electronic file storage for transmission and archiving, it is possible that XPS or a like format may supplant PDF in the future.

This completes our brief discussion of the common types of file formats and optical media designed to store electronic documents for long periods of time.