

## Digitization Reconsidered

*By Jim Lindner*

Taking a position against wholesale digitization of audio and video tape is currently an unpopular stance. As the calls ring throughout the organization to scan or digitize anything and everything in or out of sight, one who suggests caution is perceived as either a corporate anachronism or a hopeless computer phobic. The author is neither, and while there are many cases where so called "digitization" may be warranted, there are also many cases where it is a poor decision. While many arguments for digitization are propagated by vendors and corporate technocrats (who may have other agendas of their own), the reality is that there is no "one" form of digitization, just as there is no one "truth". The major benefit offered by a digital migration is the lack of multi-generational loss in making successive copies. A strategic and in depth examination of the various aspects of storing a large collection solely in the digital domain reveals that such a process may be fraught with long term risk in many areas.

This risk threatens both the life and the integrity of the material (which such a process is supposed to protect), and far outweighs the benefit of reducing the disadvantages of multi-generational loss. Indeed, what difference does the lack of multi-generation induced "noise" make if you cannot retrieve the information in the first place? This article examines just a few of the aspects and decisions involved, and takes the admittedly controversial position that a singular digitization strategy may not be the panacea hoped for at this time.

What do people mean when they talk about "digitization" anyhow? After getting up the courage to ask such a seemingly techno-ridiculous question to the local techno-guru, one usually gets the patient and patient explanation that all of the information is nicely and neatly stored as ones and zeros, invulnerable and accurate throughout time. Such an explanation is often offered as a technocratic security blanket. This familiar and comforting thought is inaccurate in the sense that it ignores the process by which the signal changes into those nice and neat ones and zeros in the first place, the process in which the signal is recorded to the media (most media is inherently analog), the process whereby the signal is reconstructed from those zeros and ones, and the many variables in-between. The reality is that we live in an analog world. The digitization process is taking a sample of that analog world and the resulting digital snapshot that we are taking at this point in the technology may be wholly inadequate in just a few years.

Anyone who has ever used a scanner to digitize a paper document can testify to the fact that there are many digitizations possible from one paper document, indeed it is often impossible to get the same exact scan twice. Some of the factors one could discuss would be the resolution of the sample or scan. What is the color temperature of the lamps during the scan? What is the amount and distribution of bits available to represent the color (black and white documents

have color too and are there enough bits to represent the depth of the blacks as well as the distribution of the bits available throughout the color space)? Is the color space compressed in any way? Are there optics in the scanner and if so what is the distortion across the field (very few lenses are perfect)? What are the errors in registration? What is the linearity and sensitivity of the array.... you get the point. One of the reasons that there are so many different scanners on the market is that each of them has different characteristics and will produce a slightly different scan. Indeed, from a purist point of view, it would be virtually impossible to get the same exact scan from a single document working at the very fullest resolution and color depth from two different serial numbers of the same exact model from a single manufacturer. A scan is a series of samples, and depending on your yardstick, decisions are made by (and for) you that relate to the accuracy of that sample. There is no one scan that can be identical to the original, the digitized version is a sample, and depending on many variables it may be a very good or a very bad series of samples depending on your definition of how good is "good enough". Even if a single scan is "good enough" what quality control systems and standards are in place to insure consistent results over many successive scans of different objects over a long period of time. Furthermore, what may be judged by some to be "good enough" by the standards of today will certainly not be good enough in the near future.

As processors and memory systems get smaller, faster, better, and cheaper, the tradeoffs that we are currently making in the digitization process today will become totally unnecessary in the future, and furthermore these tradeoffs may rob our future of information that is important for future techniques and processes. A good example of this lies in the field of video digitization where several techniques for image compression are used. These techniques are deemed necessary for many applications at the current technological moment because computer memory is too expensive and processors and busses are too slow to deal with the onslaught of uncompressed data. Although some systems can store uncompressed video, these systems are very expensive and store very short amounts of information -usually numbered in seconds, not hours. Furthermore, the compression used in video is termed "lossy" compression, meaning that it is OK throw away some of the information (this is a very different process from normal data file compression whereby all of the information is retained).

Lossy compression tries to throw out redundant or visually "unimportant" information to reduce the amount and speed of information represented in a picture. It is, however, precisely this thrown out information that is used in digital noise reduction which relies on high frequency information for signal reconstruction and improvement. Once this information is compressed and thrown away, it is no more. Manufacturers selling these systems often compare the visual quality to analog formats, when the reality is that compressed video systems are using an entirely different series of techniques to store and retrieve picture information. So how good is good enough? What level of compression and which compression algorithm is currently the best? Each of the

manufacturers will happily tell you that theirs is. If that were not bad enough, the results of those using supposedly the same compression techniques are widely different. Pictures compressed with JPEG (a very popular lossy compression technique) vary significantly between different manufacturers systems due to the specific implementation. Further, in motion JPEG used in video compression, individual frames may have different compression ratios so that individual frames content and perceived quality can vary widely depending on the content of the frame as well as the ability of the processor to keep up the ocean of information presented for it to very quickly digestion. Different compression algorithms have different artifacts, some of them are apparent in the first generation, and some of them take many generations to appear. Some artifacts affect the detail within the image, and some affect the apparent motion of the image. And everyone will agree that it is not a good idea to compress a picture that is already compressed using a lossy compression technique a process akin to cutting a pointillist picture with a matte knife into very neat squares and only keeping the average color values in each square.

One argument favoring digitization offers the theory that once the material has been digitized, it can be effortlessly and perfectly translated or migrated from one technology to the next. One article recently went so far as to suggest throwing out old equipment because digitized images can simply be migrated or translated over time to the new technologies that will be available in the future. An interesting and comforting thought, provided that there was a single migration path or technique for sampled data.... or any data, and that application software will be backward compatible forever in the future. One only needs to look at the current world of digital video. There are several competing companies offering digital encoders and decoders all translating to and from the same digital standard, and the prices of these systems can vary by the tens of thousands of dollars - similarly the picture quality can vary as well. Although there are standards in which the order of the information being transmitted is defined, how you encode and decode (or digitize) the video is NOT defined, but left to the market to determine. Going further, digital video converters between different standards (D2 to D1 for example) from different manufacturers are so different that the output actually looks different! How could this be if one could simply migrate one technology transparently to the next technology? File format translators are another huge problem, with some file formats not having enough information for other formats in which case one must "extrapolate" (or take a good guess) at what the data might be. There are currently several programs in the personal computer market that do nothing other than translate or migrate from one file format to another, an imperfect science at best due to the fact that newer file formats tend to have new features that did not exist on earlier versions of the software, and also tend to drop features that did not sell. Migration, then, is a far more complex issue than is immediately apparent.

How many times have we heard "just put it on a CD"? Exactly which CD format are we talking about, and which software is used to record and play back the

information? Is it yellow book, red book, what application is used, and what guarantee do we have that the application software will be around 50 years from now? Currently one of the biggest problems in the CD authoring market is incompatibility between different computer platforms for graphic performance. In fact, many "multi-media" computers cannot play some of the earlier CD's that were made just a few years ago, and similarly many computers cannot properly play many CD's recently authored using new software. How can we be naive enough to think that all of these systems will not change and be compatible, particularly with the knowledge that the sampling techniques used to put the information on the disk in the first place is a very quickly moving area of innovation?

Historically speaking, media failure is one of the biggest problems facing information retrieval - digital or analog. In a fit of frustration, I took the liberty of doodling on the top surface of a CD during a meeting when a corporate expert was giving a presentation on the archival advisability of storing the company's entire library on recordable CD's. When the meeting was done, I asked the individual to play back the CD, and when the machine spit out the disk (a catastrophic failure causing all of the information to be lost) he stared at the disk. When he realized what my doodles had done, he called a foul. I responded that in the real world, objects do get dropped, scratched, over heated, over humidified, and occasionally doodled upon. I further said that a catastrophic failure in any of these situations is, in my opinion, unacceptable. Media failure or loss can of course claim virtually any type of media which is why disaster planning and strategic duplication strategy is vital. This being the case, why is one media offered as a single preservation strategy, particularly in the light of the fact that CD-R was never invented with archival application in mind in the first place? In my opinion a technology that uses index tracks to map out the location of all of the information on the disk is extremely vulnerable to catastrophic loss and therefore unsuitable for an archival application.

What then is the solution? In my opinion, the solution is to realize that there is not ONE solution, but rather the recognition that any one strategy in a period of rapid technological innovation is apt to be the wrong one. The solution is to have a strategy that does not count on one technology (either digital or analog) or technique. An effective strategy is one that offers a number of different strategies depending on the application and preservation needs of the organization that owns the artifact. Multiple strategies offer much higher probability for survival, because in the case of a single loss or technological obsolescence there are other possibilities for recovery. The solution is to recognize that there is not one ultimate digitization, but that there are many possible different levels of digitization, and that each offers tradeoffs. The best single strategy is to have multiple strategies that take into account that all media is subject to failure, that migration is not guaranteed to be possible or advisable over the years, and that any single all encompassing solution that commits to a single technology is certain to fail.