

Preserving data for the future

During lunch with some old friends just before Christmas, one of them complained that she was having great trouble finding a copy of *Some Like it Hot* on video tape for her elderly aunt. “You just can’t find videos in the High Street these days,” she lamented. “It’s all DVDs.”

“Have you tried E-bay?” we all chorused. It’s amazing how often that name come up in everyday conversation, the universal panacea for those seeking the unobtainable or looking for a discreet way to dispose of unwanted gifts. Clearly, though, when you are forced to scour E-bay or the darkest corners of Amazon for your elusive product, it’s a reasonable indication that its days are numbered – and that the time has come to move on to the next wave of technology.

I have to sympathise with my friend – as a family we don’t relish life on the bleeding edge, either. We haven’t even achieved a complete conversion to CDs. My wife clings desperately to her collection of vinyl LPs from the ‘70s and ‘80s with their beautifully illustrated sleeves, but I have a sneaking feeling that it won’t always be easy to find the kit to play them on. Even DVDs, we’re told, experienced flat growth this year, as fast Internet downloads begin to consign more static media to the trash.

These changes happen in the consumer market all the time, though admittedly not as dramatically as in the multimedia business. New developments churn the market and drive the economy. You can either indulge in nostalgia and resist change or you can embrace change enthusiastically and enjoy the benefits of innovation. Most of us do a little of both, I guess, but in the end it’s our choice how quickly we change and adapt.

In the data center world, it’s much more difficult to cope with technological discontinuity. Many of our mission-critical applications originated in the days when everyone listened to music on vinyl, but in IT we don’t have the luxury option of throwing away our assets and starting over. Some of our more ancient hardware and software components seem impossible to replace (and if it ain’t broke, why try to fix it?), but are we storing up problems for those who need to manage this stuff in twenty, thirty, eighty years’ time? Back in the 1960s, who would have believed that punch cards would still be in use in the 21st Century (or, indeed, that mainframe code with two-digit date formats would not have been re-written by the Millennium!), but in Phoenixville, PA, the Cardamation Company is still maintaining card readers and helping users to retrieve data from outdated media.

Most data center managers would argue that they are busy enough solving today’s problems, without worrying about technical continuity way into the future. But part of the message contained in the recent wave of regulatory legislation is that, if we choose to store more and more information in digital form, we need to guarantee that it will be searchable and retrievable for as long as anyone needs to view it for any purpose. What good is a legally binding document if you can’t read it? If something as important as the Magna Carta or Doomsday Book were written today, would our descendants be able to appreciate its contents a thousand years from now? Bear in mind that, a few years ago when the Jubilee Underground Line was extended in London, engineers had to refer to paper plans that were a century old. These documents have no doubt been digitally preserved now, but can we guarantee that the digital format will be accessible in a hundred years’ time if another rail extension is required at that time?

Various methods have been proposed for future-proofing corporate data assets. One of the problems is the sheer quantity of data electronic data that companies are now forced to store. This is not so much of an issue in the transaction management world, perhaps, where volumes are relatively low and long-term significance is easier to judge; but in sectors such as medicine and engineering where file-sizes are growing exponentially, making sure that these vast archives are preserved for decades to come is not a trivial undertaking. And which storage media will survive into the future? Some pundits are predicting the demise of tape, and see disk as the only logical way forward; but numerous large businesses have vast tape archives in place and can envisage no cost-effective way of transferring their archived data to any other medium. After all, regulation is only one part of the picture; data centers still have to balance cost and performance as data makes its way through its natural life-cycle, and this balance will continue to influence the choice of storage media for many years to come. Tape will survive for as long as it remains more cost-effective than other alternatives.

Whatever physical media are chosen in the years ahead, the industry needs a simple standardized way for preserving data accessibility as we move from one storage generation to the next. There is some interest in extracting data elements from their original format and tagging them in an XML string, but this is likely to increase the size of data archives considerably, and of course it relies on the resource-hungry XML being around for ever, which is unlikely. For the last five years the IBM guys at Almaden, under the leadership of Raymond Lorie, have been working on a Universal Virtual Computer specifically to address this issue. The UVC works on the principle that, as new data is created, a small program P is archived alongside it which can decode the data at any time in the future and present a logical view of it. As long as we maintain a simple UVC emulator, we will always be able to run P and gain access to the data. The Almaden team already have an emulator working, though the jump from lab experiment to universally accepted standard is a very big one indeed.

However we tackle the problem of long-term preservation of data assets, it is an issue that will not go away. It's up to today's data management specialists and data center managers to ensure that corporate, government, and cultural data remains accessible to all future generations long after DVDs and iPods (and indeed DASD and tape archives) have been forgotten.