6. RESCUING DIGITAL MATERIALS

6.1 Preservation as rescue

So far this report has discussed the choices available for digital preservation very much in the sense of forward planning—following a set of logical, ordered steps in order to ensure the preservation of digital material. However, this is not always possible. What if one, or more, of these steps has been omitted? Documentation may be missing. Copying, refreshing or routine back-ups may not have been carried out. What if disaster strikes in the form of fire, flood or accident? Or if media have been allowed to deteriorate to the extent that the data they hold are impaired? How can such resources be recovered? In these cases, preservation takes the form of rescue.

Recreating lost data can be very expensive. For example, the US National Security Association has estimated that to recreate only 20 megabytes of data in a US engineering firm would cost $64,400. For some organisations, the loss of data may be catastrophic, resulting in the loss of business and eventually in closure. In certain cases, it may even be impossible to recreate the data.

Data recovery, on the other hand, is the attempt to salvage damaged media and to recapture the digital sequences that make up the information they contain. The demand for data recovery is such that there are now several professional companies specialising in data recovery techniques.

The ‘horror stories’ recounted below illustrate some of the problems that have arisen. Many organisations will at some point be faced with the challenge of recovering data from difficult (although probably less dramatic) situations. Even the best thought out strategies for digital preservation may have to include some element of data rescue—the ‘data archaeology’ carried out by BODC, described in the previous chapter, is a good example.

6.2 What can go wrong?

The factors which may contribute to data loss include:

- *Damage to, or degradation of, the media on which the data are stored*. Damage may be caused by disaster—fire, flood or accident. After the Challenger Space Shuttle disaster, for example, magnetic tapes holding valuable data were immersed in seawater for six seeks. They were recovered from off the Florida coast still within the instrumentation tape recorders, which were still intact but had broken open as a result of the impact. Three of the six recorders had been corroded by the effects of the seawater and galvanic activity between the magnesium in the tape reels and other metal components, such as the recorder case. The tapes were coated in salt deposits. The magnetic coating had disappeared in some reel segments and the tape substrate was eroded chemically in others. When an attempt was made to unwind the tapes, it was discovered that the recording side had stuck to the back of the tape, making unwinding impossible without destroying the data. Deterioration may also occur as a result of much less spectacular forces—through age, neglect, storage in unsuitable conditions, use in faulty equipment, or careless handling. Though the causes may be more mundane, the effects are similar—valuable information is rendered inaccessible.

- *Poor management, compounded by missing or weak links in the creation, storage and documentation chain*. A good example of the problems that can occur comes from the opening of the former East German data archives after German re-unification. The archives had been closed down very suddenly during the political upheaval. Machine-readable data had not been cared for in the same way as paper documents. Many files were no longer accessible and the supporting documentation was incomplete or missing. In the former German Democratic Republic (GDR) electronic data had been processed by mainframes in large centralised data processing centres, controlled by the state. All the tasks were carried out on orders from central government areas. PCs and office systems had not been introduced to government offices until shortly before unification. Specialists from the data centres often left to find new jobs, taking with them knowledge, manuals and documentation. To compound the problems some of the equipment and many of the tapes used for storage were in poor physical condition.
- *Obsolescence, where the hardware or software required to access the data are no longer available*. Even a planned transition to new software or hardware can develop into a crisis. An example here is provided by the Premier's Correspondence Unit (PCU) at the Government of Ontario, which went 'paperless' in 1993 using Megadocs software from Digital. It was also decided at this stage that the archival record would be created and supported by the creating office—the PCU—on behalf of the Archives of Ontario. Three years later, Digital announced that it would no longer support Megadocs and recommended Linkworks as a replacement. During preparations for the transition some alarming omissions and problems came to light. There had been no archiving of records, originally planned for optical disks—all records were online. The migration to new software would be extremely expensive. The PCU—the creating office—had no hardware platforms to support the Megadocs records, and Megadocs itself was now obsolete. The Archive faced the complete loss of three years of a key record series documenting a direct link between the governed and the premier.

Sometimes a combination of factors is involved. For example, poor management often goes unnoticed until a crisis brings the problems to light.

6.3 Damage to media

Magnetic media are still the most commonly used media for storing digital information. Although disks offer the most convenient storage medium, tapes remain the most widely used for mass storage. There are two types of magnetic media—hard and soft. To achieve permanent magnetism hard media must be subjected to a significant electro-magnetic field, but as a result they achieve high levels of magnetic remanence and coercivity—qualities which make them especially suited to digital data storage. Soft media need lower applied electro-magnetic fields, but have correspondingly lower remanence and coercivity.

Various materials have been used to make magnetic media over the years and research is still being carried out to refine and improve the process. Most media now consist of a fine layer of ferromagnetic materials suspended in a polymer binder, resting on a non-magnetic substrate.

All the components of magnetic media, including the particles, the binder and the substrate, are susceptible to deterioration. Oxidation and corrosion can reduce the magnetic remanence or coercivity of the particles, resulting in loss of data.

Even new media may lack reliability, and it is not uncommon to encounter new tapes which fail. Only a small number of firms make formulas for coating tapes and even fewer actually undertake the coating activity. The quality of the tape production process is crucial to the creation of reliable media, but very little information about this is available to purchasers. Some disasters have resulted from problems in the production process. BASF, for example, released millions of faulty tapes on to the market and were alerted to the problem by users who had been losing large amounts of data.

Magnetic media can deteriorate as a result of physical and chemical changes in the medium itself. The most significant threat is the breakdown of the binder. Tapes absorb water when stored in humid environments. The process of hydrolysis causes the polymer chains, which make up the binder, to disintegrate, and the binder becomes tacky, creating an adhesive build-up that makes the tape almost impossible to play—an affliction known as ‘sticky shed syndrome’ in which the magnetic media literally ‘sheds’ off from the backing substrate. The substrate, too, can deteriorate under humid conditions, causing mistracking, where the read head is unable to locate the data.

Faulty storage and mishandling will also cause deterioration. For example, if the tape is wound too loosely, air pockets will form within the pack, allowing moisture to penetrate and encouraging sticky shed syndrome. Winding too closely allows the edges of the tape to protrude beyond the pack where they will bend and tear. Horizontal storage may lead to ‘pack slip’, which will put stress on the substrate.

The increasing storage densities of magnetic tape, as well as encouraging its use for mass storage, have exacerbated the problems—there is more data to lose.

Data are stored sequentially on tape, which is transported past the read-write heads in a linear fashion. To access the information on a disk, the heads are moved from the edge of the disk toward and away from the centre along the radius. Hard disks are electro-mechanical devices, consisting of platters coated top and bottom with magnetic oxides. They are vulnerable to ‘stiction’, a combination of friction and sticking which is similar to sticky shed syndrome. If the silica based disk platters are subjected to high temperatures through excessive use or bad storage, the silica becomes sticky and the disk heads stick to the surface. Other common problems are associated with mechanical or electrical system failures.

The way the data are stored and the type of interface the drive is equipped with will affect the prospects of recovering data. Essential data are held in three areas of the disk—partition tables, the boot block and file allocation tables—all with different functions. For example, the boot record is a short program (in machine code) which issues the instruction to load the operating system into memory. It also contains information about the disk, such as the number of bytes per sector and the number sectors per cluster. Even in newer PCs, which come as ‘plug and play’, the boot record is stored in the first sector of the first track on a disk containing the active operating system, so that if it is damaged it is still possible to access the data stored after it.

Innovations to enhance access times and increase storage densities may introduce new risks of data loss. For example, the servo pattern, which controls the position of the read-write heads is normally stored on a platter with the data. However, increasing storage by scaling down disks means that the track sizes decrease, the amount of information that the servo has to return increases, and thus the size of storage area for the servo also has to increase. To address this problem IBM has introduced a No-ID format which stores the header or ID information in solid state memory rather than on the disk itself. This improves access times, but also creates a risk of information loss by severing the link between the raw data and the information about its location and meaning.

The difficulties with floppy disks mainly concern the way they are housed—in flexible and easy to damage casings, with media segments which can easily become exposed to the elements. The drive heads are more vulnerable to dust, debris or moisture and thus more susceptible to damage.

Optical media, now widely used for both storage and distribution, have problems of their own. Optical disks consist of a multi-layer sandwich and are reconstructed of polymers and metallics. Metallics are susceptible to corrosion and delamination. Polymers deform and degrade. Either may lead to data loss. Scratching the surface of an optical disk can affect the transmission of the laser beam and cause mistracking. Each type has its own shortcomings.

The least stable are magneto-optical discs. The magnetic layer is metallic and therefore subject to corrosion. They are also susceptible to temperature and humidity changes which can cause the magnetic layer to fracture. The reflective layer of CD-ROM disks is generally made from aluminium that can degrade by oxidation or corrosion. The protective overcoat can deteriorate and expose the aluminium layer to possible damage. The polycarbonate substrate is susceptible to crazing which clouds its definition. Both CD-ROM and CD-R (compact disk recordable) may be damaged by abrasions, dirt and oil residues, temperature and humidity. WORM (write once read many) disks, as with other optical media, are susceptible to changes in the optical properties of the recording layer which result in data loss.

What is the life expectancy of these media? Optical media have a longer life expectancy than magnetic media, and manufacturers make great claims for their durability. How much data loss or how many errors must occur before we can say that a particular tape or disk has reached the end of its useful life? There is no standardised test methodology for predicting life expectancy. And, as we have already seen, so much depends on conditions of storage and handling.

6.4 Recovering damaged media

What techniques are available for restoring data from damaged media? Most data can be rescued, given sufficient time and money—the value of the data must be weighed against the cost of the recovery. Many of the techniques require technical expertise. Here are some examples of what can be done.

- Sticky shed syndrome can be overcome by heat treatment. Heating the tapes for 24-72 hours at 45-55ºC can help clear the low molecular weight oils and residues from the tape surface; they are either melted back into the tape or evaporated off. However, this does not reverse the syndrome. Rather, it allows the tapes to be played again in order to copy the data to a more stable medium.
- The only way to tackle mistracking is to re-spool the tape. If the tape is left for a reasonable length of time before re-use, the distorted substrate may be reconditioned.
- While water will damage tapes over time, its effects are not immediately apparent. The absorption rate is slow and tapes can survive in clean water for days, even weeks. Tapes exposed to contaminants in the water, such as salt or mud, can be rinsed in distilled water or cleaned in soapy water.
- Drying tapes is a delicate operation and is best done naturally at room temperature.
- Disks may be cleaned and dried in the same way as tapes. The disk tracks are laid down as concentric circles, or one continuous spiral, so the disk should not be wiped in a circular motion, but from the inside out.
- Fire damage can be much more severe. If the tapes have suffered extreme temperatures, the substrate and binder may have melted, making it impossible to unwind. However, even here some recovery may be possible using thermal reconditioning processes.

New techniques for data recovery may emerge in the next few years. One example is magnetic force microscopy (MFM) which developed from research into the use of atomic force microscopes. It is possible to examine the surface of a disk or tape using a magnetic force microscope and, in the case of damaged disks, by increasing the sensitivity of the MFM, to observe the data in the damaged area. In other words, it may be possible literally to read the magnetic tracks on the disk and to use optical recognition technologies to recapture the digital sequences.

6.5 Recovering missing links in the creation, storage and documentation chain

Retrieving the raw data is insufficient on its own to recreate the digital resource which is at risk. Archivists also need information about how the data are structured, including compression and encoding, in order to reconstruct and interpret the digital sequences. When data are recovered from old media, or from media of unknown provenance, this information is often incomplete or missing altogether.

When tapes and disks were first developed for computer data storage there was considerable transparency in the layout of the data on the storage device. The 0s and 1s of the binary data being recorded were directly related to the patterns of magnetisation on the storage medium. However, responding to demands for greater storage capacity, manufacturers began to use a variety of techniques to pack more data on the storage medium with the result that stored data have lost the transparency that was evident in earlier systems.

Compression techniques allow more data to be stored, but without the decompression algorithm the raw data cannot be returned to their intended form. Compression techniques usually work separately with each data block, so the patterns in each block may differ considerably. The decoding key should be included as part of the data stream. But if no information is available on the compression method used, it cannot be assumed that decoding one pattern will necessarily work for all the others.

The basic method of recording data on a magnetic surface is known as NRZ (Non Return to Zero). This allows a 1 bit to be represented by a change in magnetic polarity and a 0 bit by no change. However, there are many different ways of encoding data, and in the absence of any documentation to identify the encoding method used, anyone hoping to restore the data will face a daunting task.

Specialist companies use a mixture of specially designed hardware and software to recover data of this kind. Very often, though, the crucial factor is experience, the ability to recognise familiar structures and to piece together the vital information from the scraps of surviving evidence. The process has been compared to deciphering enemy radio messages during the war—work which was carried out with stunning success by Alan Turing at Bletchley Park.

6.6 Overcoming obsolescence

As discussed in chapter 4, migration to new hardware and software platforms generally represents the best strategy for preserving digital information. Computer operating environments change and develop very quickly, making migration both an essential and an ongoing process. Nevertheless, there will always be cases where data collections are found which are only accessible via software and hardware which have already become obsolete. Recovering data from obsolete systems is an area of particular interest to archives, libraries, researchers and academic institutions.

A number of specialist companies provide obsolescence recovery services. Such companies generally maintain a ‘library’ of peripheral devices and can build or adapt existing platforms to read outdated data, allowing them to be transferred to new software or hardware. These services are not cheap—but they do cost less than recreating the data from scratch.

In some cases the interest and value lie, not so much in the data (or content), but in the appearance and actual workings of hardware and software. Here, the challenge is to find ways of allowing the hardware and software to function in the way originally intended, to ‘archive’ the nuts and bolts of outdated technology. The techniques include restoration, simulation and emulation.

The Computer Conservation Society, a joint venture between the Science Museum and the British Computer Society, has undertaken work in the restoration of early computers and the recreation of software necessary to run them. The work involves designing and building electronic circuits to read data for which no documentation exists, and developing software to explore the structure of the data captured.

Simulation is the attempt to imitate the functions of a piece of hardware or software. Work on the Ferranti Pegasus, an early computer, has compared the two techniques: the restoration of the original machine to working order, and the construction of a simulation to run in a windows environment on a PC-compatible machine.

Emulation imitates the internal design of hardware platforms and operating system software, allowing the original software to be run. Numerous emulators have been produced for early computer games and console games, and the technique is now also attracting interest from software companies and computer scientists.

While these techniques do not in themselves offer an effective solution to the problems posed by technological change, their efforts are not just of historical value. Experience of building emulators may help to restore valuable data that are otherwise inaccessible, and the expertise gained in restoration projects may help with the 'detective work' that is necessary in piecing together incomplete data.

The faster the pace of technological change, the more urgent is the need to find ways of overcoming obsolescence. Research is continuing to explore new techniques for data recovery. One promising new development is retargetable binary translation (RBT). A binary translator automatically translates a binary executable program from one machine running a particular operating system and using a particular file format, to another platform running a different operating system and using a different file format. It has three parts:

- The front-end works on the source code of the program on the machine which it depends on, disassembling it and converting it to a transitional format.
- The middle-end, which is independent of machine and operating system, performs the core analysis for translation, optimising the code where necessary.
- The final stage, the back-end, generates the code for the second machine, using the binary file formats of the host operating system, and is dependent on that machine. It performs this task using conventional compiler code generation techniques.

A great deal of research is being devoted to this area. The aims is to develop general, platform-independent techniques for binary translation, and to provide tools that will make software available quickly on new machines, without requiring source code or re-programming.

After this glimpse into the future, we should now return to the case histories introduced at the beginning of this section. What did they do? How did they recover their data?

6.7 Recovering the Challenger data tapes

The National Aeronautics and Space Administration (NASA) asked IBM to assist in recovering the data. They wanted to understand the mechanisms of the interlayer adhesion, to develop a process to clean and unwind the tapes, and to transfer the data on to new tapes so that NASA could piece together the reasons why the shuttle had so disastrously crashed. In the process they also wanted to find ways of avoiding this type of problem in future. The recovery team started by analysing a tape identical to that used in the shuttle in order to identify the chemical processes involved in the interlayer adhesion. They also examined the condition of the tape binder to establish whether any special handling would be necessary because of its physical deterioration. They studied three of the shuttle's recorders and the tapes they held, which contained information on engine operations, the voice recordings and the conditions of two satellites in the cargo bay.

Various processes were used to analyse the tape samples: scanning electron microscope (SEM) with energy dispersive analysis of X-rays (EDAX); X-ray fluorescence; diffraction; and chemical analysis. These techniques revealed white deposits on the edges as well as in isolated areas of the recording side; they were also found on the back side of the tape from the transferred recording side. Further analysis identified the deposits on the recording side as magnesium hydroxide. An experiment was conducted to confirm that the magnesium hydroxide was responsible for the interlayer adhesion, rather than the urethane rich binder degradation found on the back side of the tapes. Dilute aqueous acid, which dissolves magnesium hydroxide, was found to be effective in releasing the adhesion.

Having identified the chemical events which had caused the interlayer adhesion, the team had to find a method of applying the solution which would allow the fragile tapes to be unwound without further damage. They removed the eroded hub from the tapes and mounted them on a spring-loaded plastic ring. The tapes were then soaked in a tank containing a solution of aqueous acid, followed by a series of methanol washes and water rinses, and finally relubricated. After chemical treatment the tapes were unwound very carefully at a speed of 0.15m per minute.

The team realised that the tapes were very fragile and would only stand one playing to retrieve the data. They re-recorded the data in analogue mode to reduce the difficulties with signal amplification. NASA was able to read more than 90 per cent of the data from one recorder and 100 per cent from the others—allowing us, the public, to hear the voices of the astronauts immediately prior to the explosion.

The recovery team recommended that NASA no longer use magnesium as the hub material to avoid the problems of interlayer adhesion in future.

6.8 Recovering electronic records after German re-unification

Of particular interest to archivists was the Kaderdatenspeicher, or database of party functionaries, which contained personal data on 331,980 staff members of East German government agencies. At least one copy of the database is believed to have been destroyed before unification in order to protect party members. The one surviving copy was acquired by the Federal Archives.

As well as the raw data, archivists need the documentation that accompanies the digital material, in order to interpret the data correctly and place them in the relevant context. Where this documentation is missing or incomplete, as in the case of the Kaderdatenspeicher, much detective work is necessary to piece together and decipher the meaning of the data.

The volume labels, headers and first blocks of data were printed out. The volume labels and headers followed IBM format and were therefore easy to read. These data provided details of the content of each tape and an idea of the different applications of the database. At this stage a typical problem emerged: three different ways of expressing dates were found in the volume labels and headers. Many more were found in the data themselves. These included: the fairly standard dd/mm/yy; the day, month, year and decade expressed as a series of characters (e.g. V18); and bit counting, where a bit is counted for every day from a fixed date—very difficult to decipher. Similarly, different methods were used to encode the personal identification number, which every GDR citizen had. It became apparent that without further documentation on the precise structure of the files, it would be extremely difficult to decipher the exact meaning of the data.

The documentation was reconstructed piece by piece from various sources. References to the Kaderdatenspeicher were found in the paper files of the Council of Ministers which had been transferred to the Federal Archives after unification, and these records supplied the information required to decipher the data file structures and codebooks. The East German data processing centres had exchanged information using shared codebooks, and these were used to help identify codes in different data fields. It was also discovered that the data from the Kaderdatenspeicher were closely linked to other government data files, such as the database for Public Education. While these reconstructions helped archivists understand the data file structure, they still had difficulties in interpreting the data. In some cases specialist software was developed to help in this; in others they had to resort to human contact—calling in former employees for assistance. The database is now archived and copies are available for research.

This is a vivid illustration of the need for relevant documentation to support digital material. While it is possible to restore data with incomplete documentation, this is a painstaking, time-consuming—and expensive—exercise.

6.9 Rescuing the records of the Archives of Ontario

The problems faced by the Archives of Ontario were not in the same league as those in the aftermath of a disaster such as a fire or flood. Nevertheless, decisions had to be made quickly in an atmosphere of some stress and urgency, and, perhaps inevitably in such a situation, resulted in a compromise. The Archives, which had no electronic records programme and no emergency resources, was unable to undertake the preservation or migration of the records itself; nor could it take unilateral action to rescue the records.

In the event the Premier's Correspondence Unit arranged the financial and staff resources to migrate those records which dated from an important transition in political power (in 1995), which it felt were necessary for current needs—about 80,000 files. A small selection of records which predated the 1995 political transition—about 22,750 files—were printed out on paper with accompanying printouts of indexes. This was an emergency measure, taken because the resources to migrate to new software were not available, but imperfectly executed as the system was not designed for such an event. Although the records were rescued, in the sense that the data were not irretrievably lost, the opportunity for a planned, orderly transition was missed. Decisions were taken in a hurry and expediency became the priority.

The episode did serve to highlight weaknesses in the system that could be avoided in the future, and the digital stable was finally locked (even though the data had bolted). Schedules have been drawn up for archiving new records on computer-output microfilm (COM).