RISK OF LOSS OF DIGITAL DATA AND THE REASONS IT OCCURS

CASE STUDIES AND ANALYSIS

Background note

The Digital Preservation Coalition carried out a survey of its members in late 2003 to gather information about the state and needs of digital preservation. The results of the survey were discussed at a workshop in November 2003. A series of recommendations for further action were agreed. Recommendation 8 read:

> That the DPC considers establishing a log of examples of loss of digital material in order to build up a clearer picture of vulnerability and to assist in strengthening the case for action.

In agreeing that this merited further work, those attending felt that it was important to present positives. Good examples of proper management of digital assets, including provision for their long-term preservation, would be better than negative examples solely of loss. It was also important to use such examples to show how data comes to be lost, which would enable action to take suitable precautions. This is not just a question of outright loss; there are also risks to do with loss of functionality or of links, or of degradation of material. With this in mind, DPC members were subsequently invited to contribute ideas for scenarios, specific examples of data loss, and case studies illustrating the implications of inadequate management early in the lifecycle of a resource. This note contains the results of that exercise.

The DPC is indebted in particular to staff of the Arts and Humanities Data Service, the National Archives and the National Digital Archive of Datasets at the University of London Computer Centre for the examples below.

The format of this paper is:

Case studies
Lessons from the case studies
General commentary
Conclusions

Case studies

Case study 1, Newham Museum Archaeological Service.

An example of how organisational change can exacerbate risk of loss of access to information, in this case of a unique record of archaeological excavation, even where the record itself has not been totally lost.

Newham Museum Archaeological Service was closed down in 1998. Its digital archive was passed to AHDS Archaeology by the London Borough of Redbridge. The archive represents some 10 years of fieldwork and incorporates the work of other units that had previously been closed including those associated with the Passmore Edwards Museum and the Manor Valley Museum. The archive as delivered consists of about 230 floppy disks containing over 6000 files totalling over 130 Mb of data. The files were in a variety of proprietary software formats and versions some of which are now 'archaic'.

This is presented as an example of organizational failure leading to loss. It should be noted that, strictly speaking, none of the data has actually been lost. Indeed, the first priority for action by ADS was to migrate all files to the ADS file server where they are included in the general backup strategy and are hence safely preserved. However, in terms of access and use, parts of the files have become inaccessible. This might better be described as ‘information loss’, rather than ‘data loss’. Some, if not all, of the inaccessible information could be recovered but there are no resources to set about doing this. Those parts of the data which were recovered successfully are online at: http://ads.ahds.ac.uk/catalogue/projArch/newham/newham_intro.cfm

For published discussions of this case, see: Austin T, Robinson DJ & Westcott KA (2001) A digital future for our Excavated Past in Z Stancic and T Veljanovki (eds) Computing Archaeology for Understanding the Past: CAA 2000, BAR International Series 931, ArcheoPress, Oxford pp 289-296.

Case Study 2, RECAP - Rescue of Complete Archaeological Projects

This shows a case where the potential for analysis and reuse was only recognised later, by which time the original record was found to be at risk of loss and its future remains uncertain.

AHDS Archaeology and the ADS are also involved in the RECAP project, which is attempting to recover data from some of the major English Heritage sponsored excavations from the last twenty years. The RECAP project looks at 18 different projects (but this represents a small sample) which range from the archaeology of the whole of Lincoln to smaller, more compact archives. This is a mixed bag because the projects were diverse. More importantly, digital data was not seen as an important component in the project plans until after they had been completed. The problem with the RECAP data was that it had never been intended for public consumption, but was largely a by-product of the analysis and publication process. Whatever the causes of this problem, there is an additional one in that the RECAP is now missing, so there are two tasks: the first to find out where it is, the second to find out what state it is in and whether it can be recovered for access.

Case study 3, ARENA (Archaeological Records of Europe - Networked Access:

A case where obsolete software, needed to migrate old data, could not be found so only the existence of a paper record enabled the reinstatement of digital information through re-scanning.

The Terraconensis project, from the 1980s, was an archaeological landscape survey http://ads.ahds.ac.uk/catalogue/projArch/tarra_var_2003/. The record consisted of some very old databases that were updated by the depositor but became part of a larger digitisation project. This was because, in order for them to be useful, the databases needed other data sets to go with them that were in paper form. All the maps for the landscape survey were in Aldus Freehand v2 for the Apple Mac. This software required a separate package with an ‘export’ function for saving its data into other formats, and ADS was unable to find this software. In order to reinstate the data, therefore, they had to scan plans from the project publication itself. This is an example of a case where two factors combined: the first, inability to find the right version of obsolete software without expending great effort and resource in the search; the second, data loss only being avoided because the paper originals from an initial scanning exercise still remained in existence and could be re-scanned.

http://ads.ahds.ac.uk/arena/)

There is a short description of this (and other issues arising from Arena) in: Kenny J and Austin T (2004) 'Data preservation: Exploring the 'rescue' role of the Archaeology Data Service' in Content Management Focus vol 3 issue 5, 25-29

Case Study 4, UKDA/AHDS History Study 4170 - Parliamentary Poll Books of Sandwich, Kent, 1831-1868 (Cornerstone) - obsolete software

A case where the key to saving the information lay, not in obsolete software because this (with some difficulty) was available, but in hardware and the need to track down an obsolete computer with the right operating system.

This study was deposited with AHDS History in 2001, as a set of data files for an obsolete database package called Cornerstone. Cornerstone was developed by Infocom, the pioneering computer adventure game company, in the early 1980s. It was critically well received, but failed commercially, leading in large part to the failure of the company. AHDS History’s initial attempts to recover the data involved attempting to decipher the format and extract the information directly from the files, but certain details eluded them and they were not able to discover format specifications on the Web. Fortunately the depositor had been able to supply a copy of the Cornerstone software along with the data. However, in order to use this AHDS History had to locate an older computer and operating system that still supported early DOS based programs to enable them to load the software. This they managed to do. Once loaded, it was then possible both to access the data and to use Cornerstone's own export functions to convert the data to a delimited text format.

Case study 5, Unreleased UKDA/AHDS History study

A case where the stored data may survive on an obsolete disk drive, but no means has yet been found to verify this or, if it survives, to recover the data. This is despite extensive work both on hardware and software issues, and the case is unresolved

AHDS History received a 5.25 disk containing data compiled on a Shelton Instruments Sig-Net (http://www.old-computers.com/museum/computer.asp?st=1&c=810). The machine, which was launched in 1981, used the now-obsolete CP/M 2.2 operating system and stored data on a 5.25" disk drive. A faded note on the disk indicated the data had been last modified on 12th July 1988. The depositor no longer had access to the machine on which the data was created and could no longer remember which file system had been used to store the data. AHDS staff faced four problems if they were to be able to read the disk structure and extract any data files stored on the disk:

1) the operating system and original hardware were now obsolete.

2) Incompatible variants of the CP/M operating system were produced. Different implementations of the CP/M disk format make it difficult to access the data, even when tested on an actual CP/M system.

3) data was stored on 5.25" disks, which are no longer manufactured and difficult to locate.

4) The disk or file system may have deteriorated during storage, or been damaged in transit.

Two options are available in this situation - install the original operating system (either natively or via emulation), or find some method to read the 'alien' disk file system and extract the data using a current operating system. Unfortunately, neither of these methods were successful in reading the disk. The CP/M for Intel machines was a later version of the operating system, which introduced some incompatibilities and the DOS-based conversion software was unable to correctly recognise the disk structure. Based upon these results, two possible conclusions may be made:

1) the disk file system and data has been corrupted, or

2) the file system for this specific machine has not been fully documented.

AHDS History is continuing to work on this study.

Case study 6, digital records of the Schools Census

A case where an earlier migration of data had led to its corruption, and no paper documentation survived with the data to enable the lost information to be pieced together. A search for other copies of paper documentation had to be launched, and succeeded.

This case study concerns data held by the National Digital Archive of Datasets (NDAD) on behalf of The National Archives. The Schools Census is a survey of schools in England and Wales, believed to have been started in about 1946 and first recorded on computer media around 1975. The datasets transferred to NDAD are normally in the file format of Qstat, the survey software used by the Government's educational statisticians. A small number of records in one of the earliest datasets proved to be incomplete when data was extracted from this format. There was a more serious problem with the data dictionaries that gave the names of columns, explanations of their use and keys to encoded values. The data dictionaries had themselves been migrated in 1991-2 and data had been lost from them, the explanations provided for some columns having been cut short. In these and other cases column descriptions were duplicated, even though the columns clearly contained different data. Also, the dictionaries also contained unexplained abbreviations and unexplained encoding that made it impossible to define the meaning of each column precisely. It would normally be possible to overcome these deficiencies by referring to the paper documentation and copies of the original survey forms. For the very earliest 1975-9 datasets (and some later years), these documents had apparently not survived.

The solution was to engage in some 'digital archaeology' to recover the early metadata. With the help of two County record offices and a few schools, some copies of completed survey forms were located, along with instructions for completing the forms. The completed forms were matched up to the corresponding digital records, and the meanings of the columns became clear. The annual volumes of education statistics in which the results of the 1975-9 schools census were originally published gave further help, by explaining some of the abbreviations and coding schemes used in the survey. As a result, the catalogues for these datasets were brought up to the normal high level of completeness.

Two problems are identified in this case: apparent loss of data on migration, and lack of metadata information. As with case study 3 above, it was possible (though with rather more research effort needed) to track down original information still held in paper formats and use this to reinstate the loss.

For further background on the Schools Census see Peter Garrod, "The Schools' Census and Digital Archaeology", in Digital Resources for the Humanities 2001-2002: An Edited Selection of Papers, ed. Jean Anderson, Alastair Dunning and Michael Fraser (London: Office for Humanities Communication, 2003)

Case study 7, recovery from multiple copies

A case of being able to piece together a complete set of data when several copies have all become degraded.

Even where suitable procedures have been followed for a lengthy period, data is put at risk when there is a break or lapse in care, for whatever reason. That this is a very real danger is illustrated by the problems encountered in continued archiving of older datasets. In some cases these have been stored on their original magnetic tapes for decades, a preservation strategy not to be recommended. The tapes deteriorate and may become completely unreadable. At best they may be readable at a slow speed, and even then portions of the data may be lost beyond recovery. In this case, tapes were created 30 years ago, and properly and actively managed by the creators for many years. The tapes were periodically checked and rewound, and copied when necessary. Technological change meant that for the last several years of this period of storage, these provisions could no longer be maintained out as the owners no longer had 9-track tape drives. For whatever reason, no alternative strategy was implemented. The length of this period of lapse in managing the magnetic tapes is not recorded, but was not more than 7 years, possibly fewer. Thus, the problems all seem to have occurred during that time, and despite the previous period of proper and professional care. NDAD and TNA managed to recover the data, but only because three copies of each tape had been kept. They were thus able to piece together a complete set of the data by reading different segments from different copies. However, even this solution relied on actually finding one of the original 9 track open reel tape drive for reading the tapes – exactly the problem hat had caused the original failure in management. Despite its exceptionally long life as a standard interchange medium, this is a historic format and manufacture of the drives has now ceased.

Another aspect – learning from the BBC

The BBC Domesday project

The best known digital preservation case study, and one about which much has been written already, is the recent preservation of the 1986 BBC Domesday project. Carried out between 1984 and 1986, BBC Domesday commemorated the 900th anniversary of the original Domesday Book by creating a modern survey of the country. Information, in the form of text and photographs, was recorded onto two 12" videodiscs playable on a BBC Master computer connected to a special LV-ROM player. It was a huge project, done with the help of EU funding and costing around £2.5 million, a very large sum at the time.

These 12” discs were recognised as a high-profile example of the possible irretrievable loss of valuable historic data due both to possible degradation and to the lack of hardware and software to read them. Two projects were launched to ensure their safety: the first by CAMiLEON; the second jointly by the BBC, LongLife Data Ltd, ATSF and The National Archives.

Both projects have been described in great detail elsewhere (eg see the CAMiLEON website, http://www.si.umich.edu/CAMILEON/ and The National Archives website, http://www.nationalarchives.gov.uk/preservation/research/domesday.htm). CAMiLEON, used the Domesday disk (amongst other examples of obsolete technology) to test whether emulation was a viable digital preservation strategy. Its aim was to emulate, exactly, the look and performance of the original. The TNA project was based on a migration strategy but also using a degree of new technology to enhance the quality of the original material because TNA wanted to provide public access, and preferred the best possible quality of service for this.

The importance of BBC Domesday lies, first, in publicising the risk of valuable data loss to a wide public. Second, in demonstrating the technical complexity of rescue operations, and their high cost. Third, in raising issues about the philosophy of digital preservation and showing that there is at present a range of debate amongst experts as to the most appropriate philosophical approach to digital preservation. This is largely due to the newness and developmental nature of the subject, but also because different projects have different objectives.

For the average user these considerations may be of little interest; the material is preserved and is available for general public access, and the quality is higher than in the original formats. The specialists will continue to debate. But the cost has been high. This may be justified as BBC Domesday offers many lessons on approaches to and practice of digital preservation. But we cannot expect – as the above case studies show – that similar resources will be available for the great majority of digital preservation problems.

The Blue Peter example

This is not really a case study, but illustrates how great the effects of not preserving material can be, and how costly. The BBC's Blue Peter programme, in June 2004, did a piece called *'Lost and Found'* which looked into the past and future of video recording. They looked back to the early days of video recording, in the 1960s, and reminded viewers that, in the 1960s and 70s, this new equipment was widely used but often it was felt that the recorded programmes were not worth keeping so, after a couple of showings, and without any consistent policy on retention, many were wiped or thrown away. It could therefore be fairly random, which survived and which didn't. They gave the example that many of the Blue Peter films and tapes still exist from around the mid 1960s onwards.

The BBC, as part of a Europe-wide project called PRESTO, is working with European partners to develop technology to facilitate and to reduce the future costs of preserving broadcast archives, by transferring them to digital media and preserving them in that form. This is now seen as an irreplaceable part of European heritage, and there was a clear risk of the disappearance of European broadcast archives, and the extinction of 75 years of Europe's recorded historical and cultural memory. More than half of broadcast archive holdings are now ageing, and need preserving before they deteriorate beyond recovery. Equipment to play audio recordings (on vinyl and tape) - and video recordings on videotape - is now also ageing or obsolete and in short supply. Spare parts and skilled operators are also fast disappearing. Before the start of the BBC's work on this, and the PRESTO project, all broadcast archive material from the beginning of broadcasting to roughly the 1980s was seen as being at risk, and the need to do something about it was becoming more urgent each year. The need, and the costs, are great.

Blue Peter also looked at what could be done in the case of one of its programmes from December 1980 which was thought to have been lost. They showed how a BBC restoration technique could be used to put together the 'lost' programme from various surviving elements still existing on VHS tape and on film. This can be seen at: http://www.bbc.co.uk/cbbc/bluepeter/show/whatwason/factsheet_2004_06_23.shtml

Lessons from the case studies

These case studies show that a variety of actions, or inactions, can result in partial or complete data loss. While these case studies have often managed to successfully recover data so that it has not been permanently lost, this has only been achieved because of expert and often labour intensive intervention. There is often also an element of serendipity at work which clearly cannot be relied upon as a matter of course. Finally, the timeframe is critical, if these resources had been left longer, it may not have been possible to recover them, even given expert intervention and resources.

Sometimes, where data cannot be recovered in such cases, this may often be because of a resource issue, rather than a technological one. If more resource were available, steps could be taken to recover the data. This raises the closely related issue of cost-effectiveness. Even where resource is available – and often it is not – there are limits to what it is worth spending on recovering any particular set of jeopardized data. It is a matter of judgement, in such cases, how much effort and cost is justified.

The case studies also show that, while preservation problems may arise because of technology-based issues, the reason data is lost can also, quite commonly, lie in management failure rather than, or as well as, in technology failure. Such management failure can be driven by a number of factors, the most common being bad planning, unexpected change, shortage of resource, or lack of awareness or knowledge as to what to do. The absence of organisational structure, and clearly defined roles and definitions has also played a part in lack of action which might have prevented subsequent, more costly, rescue work. These case studies graphically illustrate the crucial role of the creator, but without awareness of this role and/or incentive to act, there will inevitably be many more examples of risk of potentially valuable data.

Various types of management failure can make data more vulnerable and its loss more likely. The case studies above show that these include:

Management or organisational change. For example, a body goes out of existence and its holdings pass to a successor body. The successor body may lack the means, or the inclination, or the knowledge, to take the necessary actions so loss occurs. Alternatively, one of the companies involved in creating or supplying the hardware/software may fail, making it difficult for the holder of the material to meet preservation obligations. (case studies 1 and 4)

Lack of record keeping. Information and/or metadata about data collected as part of a project is not properly recorded at the start of the project. Or, even if recorded, it is poorly stored or managed. So, over a period of years, details of the nature, importance and even location of the project data are lost and it cannot be recovered without great effort, if at all. The common result seems to be that crucial knowledge such as what systems were used to create data is lost, even if the physical medium on which the data was stored survives. (Case studies 2, 5 and 6).

Failure of arrangements to continue to manage data after its creation. If plans and resources for the management of data are not properly built into the project which creates it, then its management can be neglected subsequently, until changes and developments in software and/or hardware mean that it is inaccessible without great expenditure of resource, if at all. Even if proper management is maintained for lengthy periods, serious risk is incurred if it does finally lapse (Case studies 3, 4, 6 and 7).

The overall conclusion is that the loss, or heightened risk of loss, of digital data is usually due to a combination of circumstances involving both technical issues and management ones. Common elements are: poor forward planning; poor record keeping; a failure of custodianship at some point, perhaps due to a shift in responsibility between organisations; failures in the technical processes of custodianship.

General commentary

Perhaps the overriding point here is that it is usually cheaper to safeguard information when it is created, than to have to pay for its recovery later. The corollary is that, if this is not done, there may be cases where the costs or the difficulties of recovery exceed the value of the data recovered, so it has to be sacrificed even though it is, technically, recoverable.

Data is usually lost, or put at risk, through failures involving or arising from the technology, such as inability to read obsolete formats or degradation or corruption in the storage medium. But these risks are well known, and have been for many years. Proper planning at the time data is created will usually mean that, barring genuinely unforeseen events, they are avoided. Mostly the problems that arise are foreseeable.

If, at the time a project is planned, there is no clear way in which the future of the digital data being created can be safeguarded then those planning the project must ask whether it should go ahead, unwelcome though that may be.

The crucial issue here is the acceptance, in every such case, that the completion of a project involving creation of digital data does not occur on the day the data is all successfully created. It must look ahead to the date where the usefulness of the data, and the need to keep it, ceases and it can be disposed of. In some cases – for example, businesses keeping financial records – this time period is relatively short, is a matter of regulation or law and is clearly understood. In other cases, the time period is equally clearly understood but can be very much longer. Examples are records selected to be kept at a state archive, or publications to go in a legal deposit library, both of which, at least in theory, have to be retained in an accessible condition in perpetuity. Most cases, though, are less clear cut and it is for those creating the data and/or those funding it (where this is different) to make a judgement as to what they are planning for and why. They also need then to record what is needed in terms of preservation and how the project will cope with delivering it.

Ideally, these conclusions would be set out in a policy statement on retention for the project concerned. This policy should be part of the original project planning. It need not be a long or complicated document. As a minimum, it needs to set out clearly the view of those running the project on what needs to be kept, for how long, why, what the issues and arguments are, what the long term future management of the material should be and how it is to be attained.

Such an approach may be particularly relevant in those cases where digital data is being created using funds which are conditional on provision being made for the preservation management. Most EU project finding, and many UK national lottery grants, have such conditions attached and often require production of a digital preservation strategy as one of the conditions. It is not clear that recipients are always in a position to deliver what is required, or what help is available to help them do so. Nor is it clear that observance of such conditions is always followed up and checked.

Common though it is, it is not always management failure which causes problems. We should resist the view that, if we sort out our management, technological issues will go away. For example, at least one of the case studies – number 6 – records a loss of data arising from the migration process. This is a known risk and is being studied in great detail, for example by the Dutch government's Testbed project (http://www.digitaleduurzaamheid.nl/home.cfm).

Another common technological issue emerging from the case studies is the rapid obsolescence of both hardware and software. This emphasises the need for tools such as The National Archives' PRONOM, a resource for recording and accessing obsolete software products.

These two examples serve to show that, while it makes sense to emphasise the need to get your management in order, and while this will help avoid a lot of common problems, there are still very real technological issues on which we need to continue working.

Conclusions

Failures in technology put digital data at risk. But it can also, often, be failures in management or in custodianship which actually permit – or at least encourage - loss or damage to take place.

The seven case studies quoted here do not give enough information to allow a comprehensive analysis of the types of management or custodianship issue which can lead to data becoming at risk. They do, though, point to the clearest lesson, which is that it will invariably be cheaper and more effective to provide for long-term management at the time of creation, rather than having to struggle to do it later.

They also give enough insight for it to be clear that projects for the creation of digital data must always include, in their planning stages, a statement of policy on retention and full consideration of how the data created is to be managed throughout the life that is envisaged for it. If this cannot be done then, however regrettable, a decision is needed as to whether the project is worthwhile.

However the case studies remind us as well that there are real technological issues which need dealing with, for example problems arising from migration, or problems of rapid obsolescence.

All of this brings the focus back to the importance of life-cycle management. When we create digital data we must plan for its management and preservation, right through to the day when it is no longer needed and can properly be disposed of.