s t u d i e s

Digital Archaeology: Rescuing Neglected and Damaged Data Resources

A JISC/NPO Study
within the
Electronic Libraries (eLib) Programme
on the Preservation
of Electronic Materials

February 1999

Seamus Ross and Ann Gow

Humanities Advanced Technology
and Information Institute (HATII)
University of Glasgow
http://www.hatii.arts.gla.ac.uk/

ISBN 1 900508 51 6

Digital Archaeology: Rescuing Neglected and Damaged Data Resources was prepared as part of a programme of studies resulting from a workshop on the Long Term Preservation of Electronic Materials held at Warwick in November 1995. The programme of studies is guided by the Digital Archiving Working Group, which reports to the Management Committee of the National Preservation Office.The programme is administered by the British Library Research and Innovation Centre and funded by JISC, as part of the Electronic Libraries Programme.

Published by
Library Information Technology Centre
South Bank University
103 Borough Road
London SE1 0AA
Tel: +44 (0)171 815 7872
Fax: +44 (0)171 815 7050
URL: http://www.sbu.ac.uk/litc/

The eLib Studies series covers a wide range of topics, each title providing pertinent research into ways IT can be used to improve delivery of information through implementation of electronic library services. For a full list of titles please contact the Library Information Technology Centre.

Distributed by
TBC Distribution
South Lodge,
Gravesend Road,
Wrotham, Kent TN15 7JJ
Tel: +44 (0)1732 824700
Fax: +44 (0)1732 823829
Email: tomlinsons@easynet.co.uk

Printed by
CPC Lithographic Printers
Portsmouth

Executive summary

The brief for this project is outlined in Appendix 1. The study examines the approaches to accessing digital materials where the media has become damaged (through disaster or age) or where the hardware or software is either no longer available or unknown. The study begins by looking at the problems associated with media.

Planning for disaster recovery situations is commonplace in many organisations from businesses to higher education (e.g. Campbell 1988, Cunningham 1987, Heikkinen & Sanrkis 1996, Kahane et. al. 1988, Leduc 1991, Meater 1996, Menkus 1994, Meredith Corp 1996, Millen 1993, Neaga 1997, Robbins 1988, Rohde 1990, Stamps 1987, 'Thank…' 1996, Underwood 1997), but much less attention has been paid to data recovery. The assumption is that with good disaster planning data recovery will be, under most circumstances, unnecessary. The problem is that while attention has been paid to disaster planning and the identification of good recovery procedures the effectiveness of these tend to depend upon pre-disaster effort. This effort often never takes place. Backing up and off-site storage of backup media are good examples of activities, which although paid lip service are often not carried out rigorously. Of the 350 companies unexpectedly relocated by the World Trade Centre (NYC) bombing 150 ceased trading, many because they lost access to key business records held in electronic form (McAteer 1996, 100). More generally '43 per cent of companies which lose their data close down' ('When…' 1996, 31). The National Security Association (Washington DC) estimated that the 'cost of rebuilding just 20 megabytes of data in a US engineering firm is $ 64,400' (ibid.,). Of course even if it is possible to recreate the data it is often not possible to do it in a timely enough fashion. Less attention has been paid on the other hand to data recovery. The demand for data recovery has however promoted the development of commercial data recovery companies that specialise in addressing the post-crisis situation. Even in the technical literature there is little discussion of data recovery techniques and this is fast becoming a black box area in which the great bulk of the techniques are developed and understood only in commercially sensitive organisations.

Because of the way magnetic media are written it is very difficult to lose everything. With sufficient resources much material that most of us would expect to be lost can be recovered. Using for example a magnetic force microscope it is possible literally to read the magnetic tracks on media such as disks (Rugar, et al 1990; Saenz 1987). It might be possible to use optical image recognition technologies to recapture these digital sequences. While in its current state of development this would be an impractical way to recover data itself it does tell us much about how this material is actually recorded on the surface of media from tapes to disks and indicate future directions in data recovery.

The range of techniques involved in data recovery includes baking, chemical treatments, searching the binary structures to identify recurring patterns, and support for the reverse engineering of the content. As far as recovery is concerned we need to make a significant distinction between data recovery and data intelligibility. Essentially it may be quite feasible to recover the binary patterns from almost any piece of media, but it may not be so easy to understand what the content of those patterns actually represents. Developments in head technology will make it increasingly difficult to build a reader on the fly, especially when considering developments such as IBMs No-ID technology for writing disks and magneto-resistive heads.

Our initial understanding of the stability and life expectancy of particular types of media often depends upon the claims made by the media manufacturers themselves. These claims tend to reflect the exuberance of scientists compounded by the hype of their marketing teams. As a result it often proves difficult to take well-informed and secure decisions about technological trends and the life expectancy of new media. In the case of

the Alberta Hail project the team felt a great deal of data useful to the study of hailstorm physics and dynamics were at risk because they were stored on magnetic tape. They felt that the best way forward was to copy the data to CD-R technologies which the team perceived as a more stable medium than magnetic tape (Kochtubajda, et al 1995).[I] There is plenty of evidence that the stability of CD-R is over-rated (see below Section 1.1.6). Far from being a secure medium it is unstable and prone to degradation under all but the best storage conditions.

Hardware collection and conservation is attracting increasing attention (Keene & Swade 1994). Numerous institutions are preserving computer hardware and many of these are keeping it in working order (see Appendix 2). Emulation of both hardware and software and its role in ensuring access to digital materials is the subject of a number of investigations. The HATII team conducted a small experiment to appraise the viability of more detailed work in this area. We have described a small experiment conducted by HATII, which indicates to us that more research should be conducted in this area. Of all the techniques currently available we believe that the work in the area of binary retargetable code holds the most promise.

When some media have been identified which supposedly hold digital materials five main obstacles may inhibit their recovery.

- **Media degradation**
  This can be the result of:
  - storage under conditions of high temperatures,
  - high relative humidity during storage,
  - media coming into contact with magnetic materials,
  - disaster (e.g. lightning strikes),
  - wear as a result of excessive use, and
  - manufacturer defects;
- **Loss of functionality of access devices**
  This can be the result of such factors as:
  - technological obsolescence (e.g. devices going out of use),
  - the fact that components in mechanical devices are prone to wear out. The mass manufacturing of tape devices has resulted in their being made of less durable components, and
  - the fact that device drivers for older hardware are generally not supported by newer hardware;

I Many projects are taking similar decisions. For instance the National Sound Archive's Project Digitise opted for CD-Rs. Peter Copeland argued that the destination medium was selected because of its long shelf life, high quality, wide acceptance and distribution, flexibility, robustness, reasonable cost, and 'long uninterrupted playing-time' (1998, 128). The sensibility of this decision may be questionable.

- **Loss of manipulation capabilities**
  This is often the result of:

  - changes in hardware and operating systems which have made it impossible for applications to perform the same functions or access the same data manipulation routines (e.g. primitives, sub-routines, system libraries);

- **Loss of presentation capabilities**
  This might result from:

  - a change in the video display technologies,

  - the fact that particular application packages do not run in newer environments etc; and,

- **Weak links in the creation, storage, and documentation chain**
  This might result from:

  - a situation where it is possible to read the magnetic polarity changes and thereby recover the bits from the media itself, but then it is not feasible to interpret the data because the encoding strategy cannot be identified;

  - the inaccessibility of encrypted data because of a loss of the documentation in which the encryption key was stored; or,

  - a situation where an unusual compression algorithm was applied to the data before it was encoded and written on the media.

The data recovery company Ontrack (1996) did a study of the causes of data loss from among 50,000 of its clients. They found that the main causes were:

- hardware or system malfunction (44%) (e.g. electrical failure, head/media crash, controller failure);

- human error (32%);

- software program malfunction (14%) (e.g. corruption caused by diagnostic or repair tool, failed backups);

- viruses (7%); and,

- natural disasters (3%) (*Document Manager* 1996, 31-32).

This report examines five main topics:

- media and data recovery;

- hardware restoration and simulation, emulation, and binary retargetable code;

- case studies on data recovery;

- ways of preventing data and information loss; and,

- possible further studies in this area.

There are 7 appendices and bibliographies of both printed and electronic resources. Appendices 2 (List of preservation institutes and emulation software sites) and 3 (Data Recovery companies) will be of special interest, but it is worth noting that this is a fast changing landscape and new sites appear daily.

Information about data loss, recovery, and risk is very difficult to acquire. As part of a continuing project which HATII will be undertaking to produce a definitive study of the Post-Hoc Rescue of Digital Materials we have launched a website where users can log and access information on this topic: http://www.hatii.arts.gla.ac.uk/rescue. In the concluding section we have proposed that:

- more case histories about data loss and rescue need to be collected;
- more research needs to be conducted into the viability of the preservation of media access devices to ensure the possibility of access to a diversity of media types in the future. Even where emulation can be used to run programs and manipulate data created in other environments, devices to read the media prove much more difficult to recreate. Writing device drivers for older devices, although tricky, is far simpler;
- documentation for hardware and software although initially ubiquitous when products are first released become increasing difficult (and in some cases prove impossible) to locate over time. A concerted effort should be undertaken to collect documentation, including designs;
- more research needs to be carried out in the area of emulation;
- the use of magnetic force microscopy to recover data from magnetic media needs to be the subject of a programme of research;
- further work into the use cryptography to decode bit sequences is necessary; and,
- a media quality index needs to be developed. Some factors which might be included in any such index include: adhesion, abrasivity, durability, chemical stability, and error rates. Every piece of storage media should be marked with a quality rating.

It is also that clear that archivists, librarians, and information scientists need to extend their investigations of media and studies of its durability to the scientific journals where this material is published such as the *Journal of Applied Physics*.

Acknowledgements

Support from the British Library Research and Innovation Centre and the Joint Information Systems Committee is thankfully acknowledged. Our work was guided by the Digital Archiving Working Group appointed by the National Preservation Office and their comments on earlier drafts was very helpful.

The project was directed by Seamus Ross and Ann Gow acted as principal investigator. Our work was aided by our colleague Richard Alexander (HATII Technical Resources Co-ordinator) who assisted Ann Gow with visits to and research on data recovery companies. Gerrard Sweeney (HATII Technician) advised on the experiments with emulation and contributed to the production of Section 2.3.

This report was completed in December 1997.

Author biographies

SEAMUS ROSS is Director of Humanities Computing and Information Management at the University of Glasgow and runs the Humanities Advanced Technology and Information Institute (HATII). He teaches multimedia development and design, and cultural and heritage computing. His research includes digitial preservation studies, digitisation of primary materials, and the use of Information Communication and Technology (ICT) in the heritage sector.

ANN GOW is Resource Development Officer at HATII. She delivers a variety of undergraduate and postgraduate courses in arts computing and contributes to the creation and design of new courses. She conducts research into the design of course materials using information technology and digital resources, digitisation, and data recovery.

Contents