

**In accordance with the requirements for the
Masters of Science in Multimedia Systems**

**Department of Multimedia
Faculty of Computer Science
Trinity College Dublin
March 31st 2004**

Cover image by the author. It is a composite of the Phaistos Disk and a CD-ROM.

The Phaistos Disk is a clay disk found in Crete dating from 1,700 to 1,600 BC. Its diameter is 16 cm and it is 2.1 cm thick. There are 45 different symbols which make up the 242 hieroglyphic inscriptions. These range in a spiral pattern from centre to edge, on both sides of the clay disk. The hieroglyphic messages have yet to be translated (at least a widely accepted translation, there are many theories as to its content) despite numerous attempts.

Hellenic Ministry of Culture, "Archaeological Museum of Herakleion"

<http://www.culture.gr/2/21/211/21123m/e211wm01.html> accessed 20/03/2004

Declaration of Submission

This thesis is submitted to the University of Dublin, Trinity College, in partial fulfilment of the requirements for the degree of M.Sc. in Multimedia Systems.

I, the undersigned, declare that this work has not been previously submitted to this or any other University, and that unless already stated, it is entirely my own work.

Signed:

James Hayes

Permission to Lend or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Signed:

James Hayes

Summary

This thesis investigates the full range of issues with regard to the long-term preservation of digital information. Much of the research was carried out via scientific documents published to the Internet. Primary sources are from established, internationally recognised organisations and authors. Many organisations involved with conservation and long-term preservation of digital data were investigated. A thorough understanding of the issues involved was developed and used as a basis for investigating alternative solutions to the long-term preservation of digital information.

The broad scope of the 'digital problem' is outlined in Chapter 1 - Introduction. A thorough set of examples is examined in Chapter 2 - The Problems to emphasise the seriousness of the current situation and to illustrate the need for a working solution in the near future.

The next set of chapters investigates the range of issues impacting on long-term preservation of digital information. Chapter 3 - Storage Media Degradation shows the physical limitations of current storage media with regard to the long-term. Chapter 4 - Hardware Obsolescence investigates how the progress of hardware development impacts on concerns for long-term storage. Chapter 5 - File Format Obsolescence illustrates the process of file format obsolescence and its relevance to long-term preservation considerations.

Chapter 6 - Digital Archiving Issues outlines the many issues which will have to be addressed by any long-term preservation solution. This enables a structured approach to assessing any given solution's possible long-term effectiveness. Chapter 7 - Digital Archiving Solutions presents a variety of current commercial and theoretical solutions. The concept of creating a 'hybrid solution' from the best aspects of each is introduced.

Chapter 8 - Conclusion draws together all the material and concepts introduced thus far. The solution to the 'digital problem' is broken into two categories in order to facilitate both a theoretical and an immediate approach. The theoretical solution draws together concepts introduced in chapters 6 and 7 to present a 'hybrid solution' that allows varying degrees of archival permanence and digital authenticity for differing digital data. The immediate solution is shown to require educating the general computer user as to the true situation regarding storage of any form of digital information. An understanding of the limitations of storage media will go a long way to encouraging a stricter data management routine.

Appendices include more detailed information for the readers reference, as well as several further historic examples of documents which would not have survived quite as well had they been created or stored digitally.

Acknowledgments

Many thanks are due to Neill Farrell for his informative and enthusiastic input into this thesis.

Also thanks to my daughter Elise and my wife Lorraine who made me realise that there are many important reasons to plan for the long-term.

Table of Contents

1.	Introduction	1
1.1.	What's at Stake	2
1.2.	The Problem	3
1.3.	Considerations	5
1.4.	Solutions.....	5
1.5.	Definitions.....	5
2.	The Problems.....	7
2.1.	NASA.....	8
2.2.	The Domesday Project.....	8
2.3.	National Library of Australia	9
2.4.	German Reunification	10
2.5.	Ontrak Ltd.....	11
2.6.	Email	12
3.	Storage Media Degradation	14
3.1.	Magnetic Tape.....	15
3.2.	Magnetic Disks	17
3.3.	Optical Discs	18
3.4.	Solid State Memory.....	21
3.5.	Future Media	21
4.	Hardware Obsolescence.....	23
4.1.	Worn-out Hardware	23
4.2.	Technology Discontinued.....	23
4.3.	Near Miss	24
5.	File Format Obsolescence	27
5.1.	Official Standards and De-Facto Standards	27
5.2.	Proprietary Formats.....	28
5.3.	Market Obsolescence	29
5.4.	OS Obsolescence	29

5.5.	Hardware Obsolescence	29
5.6.	File Format Conversion	30
5.7.	Backward Compatibility	31
5.8.	Open Source Format Issues	32
6.	Digital Archiving Issues	34
6.1.	Longevity / Life Expectancy	35
6.2.	Access / Inter-operability	35
6.3.	Authenticity	35
6.4.	Reference Structures and Naming Conventions	37
6.5.	Data Redundancy / Reliability	37
6.6.	Total Cost	38
6.7.	Ease of Migration, Refreshing and or Conversion	38
6.8.	Uses as Backup and Recovery	39
6.9.	Documentation	39
7.	Digital Archiving Solutions	40
7.1.	Hybrid Solutions	40
7.2.	Digital - Analog - Digital	40
7.3.	Digital To Analog Micro-imaging	41
7.4.	RAID Systems	42
7.5.	SAN & NAS	45
7.6.	Virtualisation	46
7.7.	Grid Computing	46
7.8.	Global Grid Computing	47
7.9.	How does this tie in with Digital Archiving?	48
7.10.	Authenticity & Access through Emulation	49
8.	Conclusion	53
8.1.	A Plan for the Long-term	55
8.2.	What can we do right now?	57
	Appendix # 1 - Definitions	59
	Appendix # 2 - Abbreviations	61

Appendix # 3 - Farewell My Floppy	62
Appendix # 4 - Basic Layers of CDs and DVDs	63
Appendix # 5 - Definitions of RAID levels.....	64
Appendix # 6 - Computer storage timeline	65
Appendix # 7 - Document Storage Costs	66
Appendix # 8 - Other Historic Examples	67
Vincent Van Gogh	67
Sumerian Clay Cuneiform Tablets	67
Bibliography	68
Primary Sources.....	68
Other Important Sources.....	69
Internet Resources of Note	72

Tables

Table 1 - Summary of findings, Ontrack Data Recovery Europe Ltd	11
Table 2: Sample Generic Figures for Lifetimes of Media	17
Table 3: Disc type, read/record type, data layer, and metal layer.....	19

Images

Figure 1 – The chain of digital preservation. Image created by the author.....	3
Figure 2 – The requirements for long-term digital preservation. Image created by the author.....	5
Figure 3 - Periodic table created with a non-spatial font.....	36
Figure 4 - The same periodic table rendered with a spatial font.....	36
Figure 5 - Eastman Kodak Company's 'Digital-Analog-Digital' preservation process.....	41
Figure 6 - The image left is the original text seen at 4,500 times magnification. The image right is the same text after exposure to 300°C (570°F) air for 24 hours. Each disc is 2.2 inches in diameter and contains approximately 9,000 pages of text or images.....	42
Figure 7 – Overview of the IBM KB DIAS preservation flow path system.....	50
Figure 8 – IBM UVC DIAS preservation system.....	51

The digital heritage consists of unique resources of human knowledge and expression. ... Many of these resources have lasting value and significance, and therefore constitute a heritage that should be protected and preserved for current and future generations.

The world's digital heritage is at risk of being lost to posterity. Contributing factors include the rapid obsolescence of the hardware and software which brings it to life, uncertainties about resources, responsibility and methods for maintenance and preservation, and the lack of supportive legislation.

Unless the prevailing threats are addressed, the loss of the digital heritage will be rapid and inevitable.

From:

The United Nations Educational, Scientific and Cultural Organisation (UNESCO)
"Charter on the Preservation of Digital Heritage"

15 October 2003

1. Introduction

Long-term Digital Preservation: Continued access to digital materials, or at least to the information contained in them, indefinitely, i.e. beyond the limits of media failure or technological change.¹

From the 1940s and the 1950s onwards computing technology has rapidly developed from a highly specialised application restricted to research institutions and government bodies, into a commonplace, even household, appliance. This thesis is an investigation into the pace of current technological development, and its repercussions for the long-term with regard to stored digital information. The term 'technological development' encompasses computer hardware, computer software, data storage mediums (disks, CDs, magnetic tape, etc), networking systems (the Internet), and peripheral devices.

This thesis will investigate how fragile current digital media can be. Current digital media encompasses a wide spectrum of technologies that have yet to be rigorously tested by time and use. What are the long-term implications? Will digital data created currently survive and be accessible in successive generations?

Regardless of the many beneficial advances in hardware and storage media technologies, the heart of the issue of this thesis is the general misconception that digital data in its current incarnation is a stable format for the long-term storage of information. This thesis will show that this is simply not true.

Investigating the brief history of digital technology, from the first use of binary signals used to encode information up to today, it becomes clear that digital technology is still in the early stages of development as a new communication medium. We are in the midst of a long drawn out 'transition period', during which multiple technologies are competing for market domination and thus acceptance as 'the standard'. These technologies run the aforementioned gamut from hardware to software to storage media, and even when any one of them becomes established as a market standard, its position is quickly challenged by the next development by competing companies. A simplistic summary would be to say that computers are getting faster and storage capacities larger, but this ignores the over-riding problem of obsolescence and subsequent data loss.

¹ Definition adapted from The Digital Preservation Coalition, <http://www.dpconline.org/>, accessed 27/02/04

Obsolescence is a major issue with regard to long-term preservation of digital data. This thesis will investigate the various types of obsolescence which must be considered with regard to long-term preservation.²

1.1. What's at Stake

To understand the depth of this situation, consider what is at stake. What exactly is being threatened in this transitional period?

Records of every kind are not only stored in digital format but are 'born digital'. This means that they have been created on a computer via a software package. A brief survey would include:

- EU and all other levels of government down to Urban District Councils create masses of documents from legislation to simple email correspondence, all of which have legal requirements to be archived for varying minimum periods.
- Scientific data gathered by universities and other research bodies, covering environmental changes over time and unique events such as the Mars landings.
- Medical and Pharmaceutical records.
- Legal documents, wills, mortgages, contracts, etc.
- Taxation and employment records.
- Private correspondence, digital family photos and videos.

This list grows exponentially as each business type is considered and the unique record types accounted for. There is hardly a single business or institution in countries other than the third world which has not yet embraced digital technology in some form.

Before going any further, it must be made clear why digital documents created in the home and other 'amateur' environments must also be considered worthy of preservation considerations. Every day the general population generates what can be termed as an archaeological record for future generations. The most revealing and fascinating archaeological finds have not necessarily been those of ancient governments, but rather the rare personal correspondences, and even graffiti³ which is the very essence of anti-establishment expression. These personal documents give archaeologists an insight into the minds and personalities of past civilisations.

Therefore, a transparent system of archiving digital documents permanently (one that hides

² See Appendix # 1 - Definitions

³ John Ward-Perkins and Amanda Claridge, "Pompeii AD 79", Westerham Press, England, 1976, Page 35

The Pompeian graffiti preserved in that buried city comments on the riots of AD 59 which took place after a gladiatorial spectacle. The writer gleefully gloats "Campani you too were destroyed in the victory over the Nuceria". Visiting spectators from Nuceria were injured and killed in the riots, and Campani is a suburb of Pompeii, so perhaps some inter-neighborhood feuds were acted upon that day.

complexity) for the home user is an area worthy of development. Otherwise there is a substantial risk that a large portion of current cultural heritage will be lost and unavailable to future generations.

1.2. The Problem

Now that the broad scope of information this issue impacts on has been illustrated, consider some of the digital dangers to be encountered during this 'transition period'.

Hypothetically, in order to access a preserved digital document in the future, many issues must be overcome. What media has the document been stored on? Is the storage media still in readable condition? Is there access to the appropriate playback device, i.e. a floppy drive for a floppy disk? Is the original software that was used to create or read the document available? If the software is available to be installed, will it install properly onto the current operating system (OS)? If not, access to the appropriate OS will be required, which may also require access to the original hardware specifications on which that OS was designed to function. The software may also require specific hardware specifications. All of these issues may be seen as the chain of digital preservation, illustrated in Figure 1 below. Digital preservation is visualised here as an interconnected chain to show that the failing of any individual aspect (weak link in the chain) will render the digital document inaccessible.

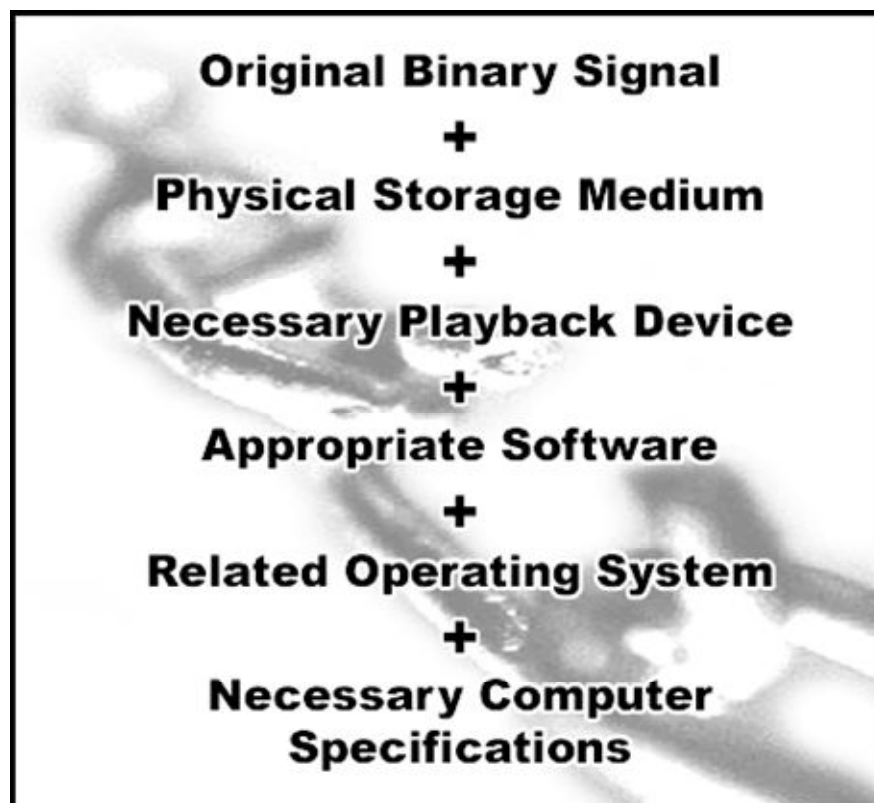


Figure 1 – The chain of digital preservation. Image created by the author.

A number of examples will be outlined in **Chapter 2 - The Problems** that show the extent to which this digital problem has arisen. This will illustrate the importance of the situation and stress the need for a working solution in the near future.

The digital storage chain of preservation can be condensed into the following major headings:

- Storage Media Degradation
- Hardware Obsolescence
- File Format Obsolescence

1.2.1. Storage Media Degradation

Under optimum storage conditions, digital storage media can theoretically survive undamaged, all of the binary data intact, for up to 75 years depending on the original storage medium used.⁴ Realistically however most businesses and homes do not operate under optimum storage conditions, but rather exist in the fluctuating temperature and humidity environments of every day life. **Chapter 3 - Storage Media Degradation** will outline the scope of the issues concerning media degradation. It will be shown conclusively that current storage solutions for digital data are not suitable for long-term storage.

1.2.2. Hardware Obsolescence

Temporary storage formats in the last 14 years alone (1990 - 2004) have ranged from 5¼ inch floppy disks, 3½ inch floppy, 100MB Zip, 250MB Zip, CDR, CD-RW, to DVDs and Flash Cards, as well as countless other competing formats of 'superdisks'⁵. This progression of storage media formats has been paralleled by the hardware necessary to access these storage units. **Chapter 4 - Hardware Obsolescence** will show that this is a separate but equally important issue to the long-term preservation of digital information.

1.2.3. File Format Obsolescence

File format obsolescence becomes an issue when stored digital data becomes inaccessible because the software necessary to interpret or decode the digital signal is no longer available. **Chapter 5 - File Format Obsolescence** will outline the many various causes of file format obsolescence and demonstrate the breadth of this issue.

⁴ See Table 2: Sample Generic Figures for Lifetimes of Media page 17 of this document.

⁵ For example the now obsolete (as of March 2003) Imation SuperDisk™ 120MB Diskettes. http://www.imation.com/en_US/product.jhtml?Id=IM_FAM122 accessed 10/03/2004

1.3. Considerations

In order to proceed effectively towards an appropriate solution, the full scope of requirements for long-term preservation needs to be outlined. *Chapter 6 - Digital Archiving Issues* will expand on the essential components of a fully capable long-term preservation solution. Each of the necessary aspects illustrated in Figure 2 below will be investigated and explained.

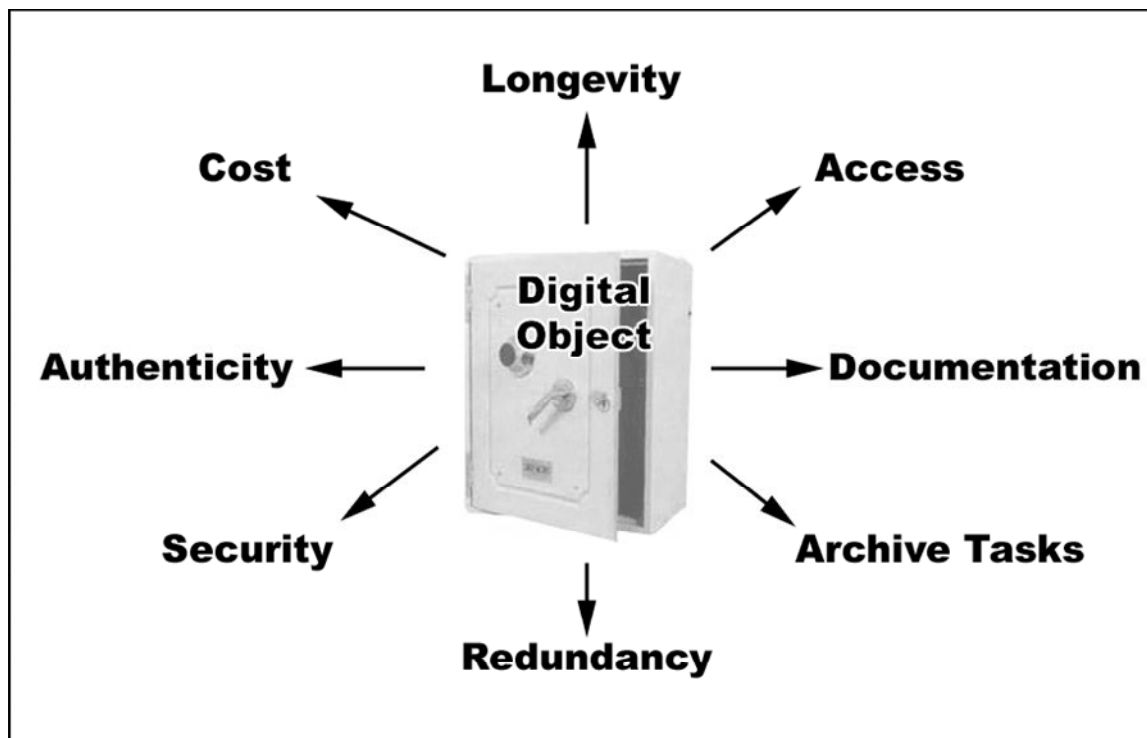


Figure 2 – The requirements for long-term digital preservation. Image created by the author.

1.4. Solutions

There are organisations, from government research bodies to private companies, who are looking for solutions to this growing problem of digital preservation. The range of posited solutions, sometimes strange and ironic, will be investigated in *Chapter 7 - Digital Archiving Solutions*. An approach to combining solutions into a 'hybrid solution' will also be posited.

1.5. Definitions

There are terms involved with long-term digital preservation, which overlap and need to be clarified before proceeding, as they are not interchangeable.

Media Degradation refers to the actual object that the digital signal is stored on, such as the floppy disk or magnetic tape. Media degradation takes place as a result of many factors, which will be outlined in *Chapter 3 - Storage Media Degradation*.

Data Migration refers to the transfer of digital signals from one storage media type to another, for example from magnetic tape to CD.

Data Refreshing involves transferring digital signals to an identical but newer storage media, i.e. tape to tape, CD to CD, etc. Under ideal conditions digital documents should be 'refreshed' to newer storage media on a regular basis to avoid data loss due to degradation of the storage media, through age or other accidental impacts. This involves vigilance and planning.

Data Conversion involves changing the digital file format to another format, such as converting from a Word file to a PDF document. Conversion almost always involves the loss or corruption of information. Again, this requires vigilance and planning. This is a much more involved task than archiving a binary string, and through the conversion process new problems may arise. Will the original document formatting or content be preserved, or will the software 'improvements' wreak unpredictable changes on the document? These issues will be covered in **Chapter 5 - File Format Obsolescence**.

Archiving in the context of this thesis does not only refer to the traditional, institutional meaning of the word. The archiving of digital data is an action which anyone working with computers needs to consider as soon as they begin to create unique and possibly valuable (historically, financially, legally, etc.) documents. Archiving refers to the entire range of considerations necessary to successfully preserving digital information in the long-term. Archiving issues will be introduced in detail in **Chapter 6 - Digital Archiving Issues**.

'Born Digital' refers to documents that are created on the computer, as opposed to ones that have been digitised by being scanned. This phrase is common to the many documents used as reference for this thesis, so no single credit can be attributed. The website "Word Spy"⁶ cites the earliest use in the article by Marcia Stepanek, "From digits to dust," Business Week, April 20, 1998

⁶ The Word Spy website, <http://www.wordspy.com/words/born-digital.asp>, accessed 19/03/2004

2. The Problems

This chapter will investigate examples of the 'digital problem' to illustrate the severity of the current situation. The diversity of situations in which problems with preserving digital data is encountered will highlight the urgent need for a working solution to this problem.

The reports started appearing in the press in the late 1990's.

- Magnetic tape containing 1970's era satellite photo survey data of the Brazilian Amazon cannot be accessed to establish deforestation trends.
- Twenty percent of the data collected during the 1976 Viking Mars mission can no longer be read.
- In the State of Oregon, the primary database of people with disabilities vanished.
- Some POW and MIA records and casualty counts from the Vietnam War can no longer be read.
- At Pennsylvania State University, all but 14 of some 3,000 computer files containing student records and school history are no longer accessible.

The "reports" above are cited in many reports and research papers verbatim⁷, but locating the original sources (aside from NASA) has proved difficult, and as such they should be viewed as modern folklore based on actual events until proven otherwise. It may be that the organisations involved are hesitant to admit to such large data losses, or that they have since successfully implemented data recovery schemes.

⁷ Documents such as "Digital Insurance For Information At Risk", "Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative", and the website for American Microlmaging @ <http://www.amibusiness.com>. Some of these documents credit Marcia Stepanek's article "Data Storage: From Digits to Dust", Business Week 20/04/1998, but she doesn't list her sources.

Ross and Gow in their publication “Digital Archaeology”⁸ support the notion that companies are hesitant to admit to losing data:

More case histories about data loss and rescue need to be collected. Good case histories are hard to find. In most instances organisations are embarrassed by their failures. We believe that more concentration of oil exploration firms might produce suitable case studies.⁹

2.1. NASA

NASA is one of the few organisations willing to speak publicly about their own challenges dealing with obsolete digital data storage systems. In the mid-1980s a joint effort was undertaken by NASA and the National Oceanic and Atmospheric Administration to recover several years’ worth of satellite data stored on magnetic tape. It cost them over US\$500,000 and they lost a large chunk of data anyway.¹⁰ The sheer volume of scientific data captured by NASA on a daily basis from the many concurrent projects - satellites, space shuttles, interplanetary missions, the Hubble telescope, etc - make data management and storage a monumental task.

According to Elaine Dobinson, Planetary Data System project manager for NASA’s Jet Propulsion Laboratory¹¹, all past mission data stored on magnetic media or paper is being transferred to CDs or DVDs. They recognise that there is a need to continually refresh media (to avoid media degradation), and that they will face data migration issues again the next time they upgrade their system to another technology.

2.2. The Domesday Project

William the Conqueror’s Domesday Book was compiled in 1085 and has lasted just over 919 years (as of 2004). The BBC’s 1986 Domesday project, involving hundreds of thousands of schoolchildren across the UK and released on videodisc, was unreadable after 15 years when the videodisc drive became obsolete.¹² Andy Finney is managing preservation work on the original Domesday videotapes for the UK Public Record Office in conjunction with the BBC. They have produced archival video masters of the videodisc contents that, according to Finney “have, as far

⁸ Seamus Ross and Ann Gow. “Digital Archaeology: Rescuing Neglected and Damaged Data Resources”, A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials, Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, February 1999, <http://www.hatii.arts.gla.ac.uk/>, accessed 23/02/04

⁹ Seamus Ross and Ann Gow. “Digital Archaeology”

¹⁰ Kridler, Chris. “Digital history is vanishing”, Florida Today, February 25, 2001. <http://www.floridatoday.com/news/people/stories/2001/feb/peo022501a.html>

¹¹ Kridler, Chris. “Digital history is vanishing”

¹² Schofield, Jack. “Digital dark age looms”, The Guardian, January 9, 2003.

as we can predict, a reasonable shelf-life using broadcast video technology.”¹³

In mid January 2003, the original 1-inch C format analogue videotapes ... were copied onto D3. D3 is an uncompressed PAL (composite) broadcast videotape format and is, for the project, an intermediate format enabling the PAL video signal to be moved from somewhere where it can be carefully played from the original, and potentially fragile, 1-inch tapes to somewhere where the PAL can be carefully decoded into component¹⁴ form for preservation. Ironically D3 is an obsolescent format in a component world.¹⁵

The BBC's 1986 Domesday project is not the only contemporary example of the many problems encountered with digital technological obsolescence.

2.3. National Library of Australia

Deborah Woodyard at the National Library of Australia carried out a trial transfer of data from floppy discs to CD-R in preparation for a special meeting of publishers and State Libraries in 1997¹⁶. Her process and findings are condensed below as a concise example of the problems encountered in a 'simple' transfer process.

The National Library of Australia has been collecting electronic publications in various physical formats since 1983. Initially 64 items were selected from over 1,439 as a representational sample of material contained on floppy disks. In order that the test would reflect the scope of the problem, Woodyard ensured that the sample publications ranged across the following variables.

- DOS vs. Mac operating systems
- various hardware and software requirements
- 5¼ inch and 3½ inch disks
- multiple disks vs. single disks
- varying ages up to 12 years old
- portion of publication on floppy disk (whole document, supplementary material, set-up software)

¹³ Finney, Andy. The Domesday Project website, <http://www.atsf.co.uk/dottext/domesday.html> accessed 02/03/04

¹⁴ The term 'component' in relation to digital technology can refer to hardware, software or data. A component is a fully developed unit which can be re-used in varying situations. See Appendix # 1 - Definitions for more details.

¹⁵ Finney, Andy. The Domesday Project website

¹⁶ Woodyard, Deborah. "Farewell My Floppy: A strategy for migration of digital information", National Library of Australia, 1997 <http://www.nla.gov.au/nla/staffpaper/valadw.html> accessed 2/23/2004

Of the 64 items that Woodyard attempted to copy from floppy disk to CD-R, 38 disks could not be used due to hardware or software incompatibilities, 3 disks were faulty or damaged, and 1 disk was discovered to be blank. By the end of the 12-step process Woodyard had copied data from 40 of the disks, but only 22 of them would run from the CD-R due to hardware or software incompatibilities. See Appendix # 3 - Farewell My Floppy for a more detailed breakdown of the process.

Deborah Woodyard summed up her findings succinctly when she stated that:

If we cannot find, operate and transfer this kind of material easily, it may in effect be useless even while remaining in a good physical condition.¹⁷

2.4. German Reunification

Deborah Woodyard's trial transfer took place as a self-imposed experiment, and produced relevant findings. However the issues of digital data archiving are rarely encountered under such controlled conditions.

One case study outlined in "Digital Archaeology: Rescuing neglected and damaged data resources"¹⁸ involved German reunification. After the reunification of Eastern and Western Germany, western archivists took over responsibility for former East German data archives. Digital information in these archives was badly organised, software and hardware documentation was missing, and the knowledge necessary to run certain systems was available only in the minds of former staff members. Significant amounts of data were lost due to low quality storage media and negative storage environments.

From the data that survived, archivists targeted the personal data of 331,980 staff members of the East German Government as being of historical significance. The team involved with this process had to start by identifying and printing out the volume labels, headers and initial data blocks of files (zeroes and ones onto paper), in an attempt to understand which file format may have been used originally to create and read the files.

¹⁷ Woodyard, Deborah. "Farewell My Floppy"

¹⁸ Seamus Ross and Ann Gow. "Digital Archaeology"

A range of specialised software was developed to reconstruct the file structures, to address problems with date formats, and to decipher binary sequences. It still proved necessary to employ former staff from the GDR archives to identify certain compression algorithms and other encoding standards that the Federal Archives team were not able to interpret.¹⁹

With time and expertise the archive team was able to reconstruct data from unknown storage media and software file formats, but it was a long and costly process.

Ross and Gow outline the main causes of data loss as media degradation, loss of functionality of access devices, loss of software manipulation capabilities due to hardware or operating system changes, changes in presentation capabilities, and weak links in the creation, storage and documentation chain²⁰. The weak links are mainly the result of loss of documentation about encoding, encryption and compression algorithms, as found in the German reunification case study.

2.5. Ontrak Ltd.

The UK based data-recovery company Ontrak Ltd. maintains statistics of the causes of data loss from among over 50,000 of its clients²¹. The simple fact that they could poll 50,000 victims of data loss who were prepared to pay for recovery indicates the scale of the growing problem of archiving and accessing digital data. Their findings are summarised in the table below. These findings may be biased, as firms that lose data due to natural disasters tend to suffer other extensive damage as well, and therefore may not be in a financial position to afford to have data recovered.

%	Cause
56%	Hardware or system malfunction (e.g. electrical failure, head/media crash, controller failure)
26%	Human error
9%	Software program malfunction (e.g. corruption caused by diagnostic or repair tool, failed backups)
4%	Computer viruses
2%	Natural disasters

Table 1 - Summary of findings, Ontrack Data Recovery Europe Ltd²²

¹⁹ Seamus Ross and Ann Gow. "Digital Archaeology", page 42

²⁰ Seamus Ross and Ann Gow. "Digital Archaeology", Executive Summary pages IV - V

²¹ Seamus Ross and Ann Gow. "Digital Archaeology", Executive Summary page V

²² Ontrack Data Recovery Europe Ltd, "Understanding Data Loss", <http://www.ontrack.com/understandingdataloss/> accessed 23/02/2004 Note: Figures are correct as of 23/02/2004 but only add up to 97%. Nominal amounts may have been excluded.

2.6. Email

Up until very recently historians relied on personal correspondence (letters) to reconstruct a profile of historic figures and events. These letters may have been on hotel stationary, handmade paper, velum, or even clay tablets, but regardless of physical form the content is for the most part immediately accessible (setting aside issues of translating ancient languages) just by looking at the document. Current historians and those in the near future face the unenviable task of accessing the currently popular form of correspondence in the developed world, email.

In the book “White House E-Mail”²³, edited by Tom Blanton, it is revealed that email documents from the Reagan, Bush and Clinton U.S. presidential tenures survive today only through the actions of a six-year lawsuit brought by the National Security Archive, supported by historians, librarians and public interest lawyers. In November 1986 John Poindexter and Oliver North deleted over 5,000 email documents as the Iran-contra scandal broke. Digital data back-up tapes were regularly recycled at the White House rather than being archived, and in January 1989, as George Bush took over from Ronald Reagan, it required a U.S. District Court temporary restraining order to prevent the destruction of the backup tapes for the Reagan presidential tenure. The same process was repeated for both the Bush and Clinton administration, and it wasn’t until February 1995 that U.S. District Judge Charles B. Richey settled the issue by declaring that

No one, not even a President, is above the law.²⁴

The attempted destruction of White House documents comes as no surprise, and shredding paper or deleting electronic signals are equivalent actions. What these events highlight however is the importance that email has taken on as an officially accepted form of correspondence, and in many cases involving correspondence of historic significance. This places email squarely within the remit of any long-term archival solution.

²³ “White House E-Mail: The top-secret computer messages the Reagan / Bush White House tried to destroy”, edited by Tom Blanton, National Security Archives, Washington, 1995.

²⁴ “White House E-Mail” edited by Tom Blanton

... email is a good example of a 'transient' means of digital communication. When the average employee is asked by the ICT²⁵ department to empty his mailbox because the limit has been reached, he faithfully does so by clicking a button, without realising that certain emails really should be preserved.²⁶

As the above quote from the working report of the Dutch Government's Digital Preservation Testbed reveals, an apparently simple digital file like email comes with its own preservation challenges. The report goes on to outline the hidden metadata in the email headers, information such as the time and date the message was sent, that will be lost if a preservation strategy only involves printing important emails to paper. Email is a way of communicating digital messages which is not platform or software dependent, yet it still encounters the same problems of proprietary file format obsolescence.

... although the email transmission file was developed expressly for interoperability, the email message is usually stored in proprietary format such as *.msg for Outlook messages or *.view for Novell GroupWise messages ... If such a message is simply stored as a *.txt or *.html file, it is unlikely that all of the essential transmission data needed for long-term preservation will be saved.²⁷

Although many email documents may have long-term historic significance, the Testbed report confirms the broad opinion that email is somehow inherently transitory and personal in nature, and therefore not subject to the normal rules of any organisation's archival policy. This notion is reinforced by the fact that the National Security Archive had to take legal action against the Whitehouse to preserve emails of historic significance. This notion of email being inherently transitory adds one more layer of difficulty for managers approaching the task of digital archiving.

It should be obvious from the examples cited in this chapter that the chain of digital preservation is very weak indeed. In the following chapters the chain of digital preservation (as illustrated in the introduction) will be investigated in detail to determine what exactly needs to be addressed with regards to long-term preservation of digital information.

²⁵ See Appendix # 2 - Abbreviations

²⁶ "Digital Preservation Testbed: From digital volatility to digital permanence - Preserving email", The Hague, April 2003.

²⁷ "Digital Preservation Testbed", The Hague

3. Storage Media Degradation

Preserving the media on which information is electronically recorded is now well understood to be a relatively short-term and partial solution to the general problem of preserving digital information. Even if the media could be physically well-preserved, rapid changes in the means of recording, in the formats for storage, and in the software for use threaten to render the life of information in the digital age as, to borrow a phrase from another arena of discourse on civil society, 'nasty, brutish and short'.²⁸

This chapter investigates examples of common data storage media currently in use and discusses the issues affecting their suitability as long-term storage solutions.

Digital technology hardware developments lead to data storage devices and media (i.e. disks) becoming redundant, which in turn leads to the inability to access data stored on these media²⁹. Added to this problem of obsolescence is the general misconception that magnetic and optical storage media to date are long-term storage solutions. The following discussion will reveal how the physical makeup of the storage media combines with atmospheric and other conditions to result in unpredictably short life spans.

Before progressing any further, it should be noted that hardware and storage media developments are not intrinsically negative. Advances in hardware capabilities over the last decade have led to tremendous leaps in real time 3D rendering capabilities. The processing necessary for generating real-time 3D animations (such as in games) is hardwired into 3D graphics cards which take the processing workload away from the main CPU allowing it to get on with other work such as managing the game logic.³⁰

Storage media developments have similarly led to viable solutions for short-term storage of very large new-media files such as audio and video. Up until the 1990s the standard removable storage for most PCs was the 3½ inch floppy which is limited to a 1.4MB capacity, which is no

²⁸ Waters, Don. "Some Considerations on the Archiving of Digital Information", Yale University Library, January 1995; <http://www.ifla.org/documents/libraries/net/waters1.htm>

²⁹ See Chapter 4 - Hardware Obsolescence.

³⁰ A good article on how 3D graphics cards accelerate 3D rendering is at Howstuffworks.com, <http://computer.howstuffworks.com/3dgraphics8.htm> accessed 26/04/2004

where near suitable for even short sound files.³¹

Regardless of the many beneficial advances in hardware and storage media technologies, the heart of the issue of this thesis is the general misconception that digital storage media in its current incarnation is a stable format for the long-term storage of information.

In his presentation at the 2nd National Preservation Office Conference in Brisbane in 1995 Ross Harvey³² clearly demonstrated that preserving an artefact such as a magnetic tape or a CD-ROM is not a viable option in the long-term. The only true consideration should be the preservation of the digital object itself, regardless of storage media, original software & original hardware platforms. This argument gains validity when the physical properties of the storage media themselves are investigated to gauge the longevity of the storage media.

3.1. Magnetic Tape

Ross Harvey outlines the physical properties of magnetic media in intricate detail; the information in following section is based mainly on his findings.

Magnetic tape, be it video or audio or data storage tape, is composed of a series of discrete layers. Magnetic particles are suspended within a polymer binder in the topcoat, the alignment of these particles directly relates to the data stored. The polymer binder is adhered to a base or substrate layer. Other substances are also added, such as lubricants and head cleaning agents.

The polymer binder is the main weak link in magnetic storage media as it is easily affected by hydrolysis, the process whereby moisture in the air causes polymer linkages in the binder to break. As the polymer chains break they become tacky. This leads to 'sticky tape' or 'sticky shed syndrome', causing the binder to stick to the playback and recorder heads, and the magnetic material 'sheds' off the substrate. The other problem is lubricant loss, which leads to increased friction and further problems of 'sticky tape'. Lubricant loss is a function of time, regardless of whether the tape is played or not.

The magnetic particles (or pigments) within the polymer binder are the next consideration. The most stable are composed of iron oxide and cobalt-modified iron oxide. High quality tapes however use metal particulate (MP) and chromium oxide (CrO₂) pigments because they can record higher frequencies and allow higher signal outputs. Although MP and CrO₂ have inferior signal stability over time compared to iron oxide and cobalt-modified iron, their deterioration can

³¹ See Appendix # 6 - Computer storage timeline

³² Harvey, Ross, 1995 NPO Conference Paper, National Library of Australia website, <http://www.nla.gov.au/niac/meetings/npo95rh.html#roth> accessed 27/01/2004

be delayed by storage at low temperatures.³³

The substrate is usually composed of polyester film, which is chemically stable. Problems with the substrate layer arise via mechanical stresses on the tape, caused by fluctuations in temperature and humidity. These can be reduced to a large degree through careful climate control of storage areas, and through careful handling of the tapes themselves. Tapes should be 'acclimatised' before playback, and the playback hardware must be kept immaculately clean as well as in good working condition.

To summarise, Harvey proposes that the main considerations for the storage of magnetic media are:

- Careful handling and packaging.
- The quality of the climatic storage conditions - low temperature and humidity.
- The number of times the tape is accessed.
- The composition quality of the tape.
- The future availability of the playback hardware.

Binder hydrolysis is the key degradation issue (amongst several) and as such humidity and temperature control is of primary concern, below 20°C and below 40% humidity as recommended by the research of Van Bogart³⁴. Although the temperature requirements might seem reasonable in the Irish context, the humidity levels are hard to control in anything short of an archival environment. In the broader scope of the globe, there are many geographic locations where meeting these requirements would be challenging. According to Van Bogart's findings a data tape, which might last up to 10 years in Melbourne Australia at 25°C and 50% relative humidity, would only last 2 years in Singapore at 30°C and 80% relative humidity.³⁵

The Digital Preservation Coalition's Handbook³⁶ lays out an even bleaker set of statistics. These findings assume no or very infrequent access to the storage media, and also that there are no other contaminants, u-v light and strong magnetic fields present.

³³ The Digital Preservation Coalition Online Handbook, <http://www.dpconline.org/graphics/medfor/media.html> accessed 27/02/2004

³⁴ Van Bogart, John. "Magnetic Tape and Handling: A Guide for Libraries and Archives" Washington DC: The Commission on Preservation and Access and National Media Laboratory, 1995; also available @ http://www.nml.org/resources/misc/commission_report/contents.html

³⁵ Van Bogart, John. "Magnetic Tape and Handling"

³⁶ The Digital Preservation Coalition, <http://www.dpconline.org/>, accessed 27/02/04

Device	25RH 10°C	30RH 15°C	40RH 20°C	50RH 25°C	50RH 28°C
D3 magnetic tape	50 years	25 years	15 years	3 years	1 year
DLT magnetic tape cartridge	75 years	40 years	15 years	3 years	1 year
CD/DVD	75 years	40 years	20 years	10 years	2 years
CD-ROM	30 years	15 years	3 years	9 months	3 months

Table 2: Sample Generic Figures for Lifetimes of Media³⁷

Van Bogart in email conversation with Harvey raises the issue that in data tapes such occurrences as signal dropouts caused by dust or debris, or mistracking caused by inadequate acclimatisation, cannot be compensated for by computers as they are by the human brain in relation to audio or video tapes. Our ears and eyes can deal with a slight hiss, pop, or annoying visual band, but missing or corrupted signals can lead to entire data files being unreadable.³⁸

However, as Vic Booyesen points out in his September 2002 paper "Information and the Archiving thereof",

The price of automated tape storage per megabyte purchased ranged from one-tenth to one-fiftieth the equivalent price for disk storage by the year 2000. Though tape may not be the optimal removable storage media forever, it is today and for the foreseeable future.³⁹

3.2. Magnetic Disks

The magnetic layer in magnetic discs (floppies) is composed in a similar manner to magnetic tape; essentially it is a polymer with a magnetic particle coating on a stiff polyester substrate.⁴⁰

Magnetic discs are highly durable but are susceptible to the same (destructive) forces as magnetic tapes, mainly humidity, heat, age of the disc, condition of the playback device, and care in handling the disc.

The drive heads of floppy disc drives are easily exposed to the elements (e.g. moisture, dust and other debris) and as such are susceptible to damage. Another issue is the alignment of read-write heads, as there is no feedback mechanism from the disc concerning whether or not the heads are properly positioned over the data tracks.⁴¹

³⁷ The Digital Preservation Coalition, Media and Formats section of the online Handbook

³⁸ Harvey, Ross. 1995 NPO Conference Paper, References 10.

³⁹ Booyesen, Vic. "Information and the Archiving thereof", Enterprise Storage Comparex Africa (Pty) Ltd., 2003, page 18. Available at the Saice IT website http://www.saice-it.co.za/pdf/24th/Vic_Booyesen_Paper_Information_And_The_Archiving.pdf accessed 09/03/2004

⁴⁰ Accurite Technologies, "Floppy Disk Drive Primer", <http://www accurite.com/FloppyPrimer.html> accessed 23/02/2004

⁴¹ Accurite Technologies, "Floppy Disk Drive Primer"

3.3. Optical Discs

Fred Byers outlines the important physical characteristics of optical discs in his informative publication "Care and Handling of CDs and DVDs - A Guide for Librarians and Archivists"⁴². The following section is based mainly on his work.

CDs and DVDs are optical media, which use laser light for data retrieval. The disc drive focuses a laser light beam into the CD or DVD to read bits (data) imprinted in the disc. Recordable and re-writable discs can also be 'burned' with new data bits via the same laser beam.

CD-ROM and DVD-ROM discs are read only discs. CD-R, DVD-R and DVD+R discs are write-once discs, i.e. recordable but not erasable. CD-RW, DVD-RW and DVD+RW discs (re-writeable) are discs from which data can be erased and new data can be recorded in the same location on the disc. DVD-RAM discs are also re-writable discs formatted for random access, similar to a computer hard drive.

CDs and DVDs are manufactured differently but consist of the same basic materials and layers. A polycarbonate substrate makes up the majority of the disc and is transparent so that the laser may read through it to the data layer. All CDs and DVDs contain data layers and metal layers, but they are composed differently depending on the type of disc, i.e. read-only, read and write-only, or read and re-writable. CDs also include a thin protective lacquer layer onto which an information label can be applied.

DVDs include a central adhesive layer that binds the two 'wafers' or halves together. Double-sided double-layered DVD-ROM discs include extra adhesive layers that bind semi-reflective and fully reflective metal layers (four in total). Single-sided DVDs can have an upper label, but not double-sided DVDs as the label would obstruct the path of the laser.

See Appendix # 4 - Basic Layers of CDs and DVDs for a full breakdown of layer order for various disc types. See Table 3 on the following page for a summary of the data and metal layer components.

⁴² Byers, Fred R. "Care and Handling of CDs and DVDs - A Guide for Librarians and Archivists", NIST Special Publication 500-252, NIST & CLIR, USA, 2003

CD-	DVD-	Type	Data Layer	Metal Layer
CD-ROM Audio/Video and PC use	DVD-ROM Video/Audio and PC use	Read Only	Moulded	Aluminium (also silicon, gold, or silver in double layered DVDs)
CD-R	DVD-R DVD+R	Recordable (Write once only)	Organic dye	Gold, silver or silver alloy
CD-RW	DVD-RW DVD+RW DVD+RAM	Re-writable (Write, erase, and re-write)	Phase-changing metal alloy film	Aluminium

Table 3: Disc type, read/record type, data layer, and metal layer.⁴³

The data layer in CD-ROM and DVD-ROM discs is pressed into the polycarbonate substrate surface. The data itself is recorded as minute pits and lands representing the zeros and ones sequence. This is then sprayed with a fine layer of reflecting metal, which the laser uses to the read data by detecting the changes in the light reflected off this surface.

The data layer for recordable discs is composed of an organic-dye sandwiched between the polycarbonate substrate and the metal reflecting layer. Although different manufacturers use different dyes, they are all photosensitive. Bits of data are burnt into the dye by the laser beam.

This dye degrades over time, eventually making the disc unreadable.⁴⁴

The data layer for re-writable CDs and DVDs is composed of a phase-changing metal alloy film, sandwiched between the polycarbonate substrate and the metal reflecting layer. The laser writes or erases data by heating this layer to specific temperatures to change minute spots from the original crystalline state to an amorphous state (write) and back (erase).

The lifespan of optical discs depends on several factors (as shown with magnetic tape) including:

- Type of disc (CD-R, CD-RW, DVD-ROM, etc) as this dictates how susceptible it will be to environmental and other factors.
- Quality of production.
- Initial conditions before recording data
- Handling and maintenance
- Environmental conditions (temperature, relative humidity, and exposure to light)

A few of the specific conditions which negatively affect the lifespan of CDs and DVDs are expanded on next in order to more clearly understand the many influences at work degrading data until it is unreadable.

⁴³ Byers, Fred R. "Care and Handling of CDs and DVDs" Page 5

⁴⁴ Byers, Fred R. "Care and Handling of CDs and DVDs" Page 8

When the metal layer is composed of aluminium it is susceptible to oxidation which diminishes its reflectivity. This makes the disc unreadable and is sometimes referred to as 'disc rot'⁴⁵. This may come about as a result of low quality production methods. Prolonged humid environments or direct contact with liquids can also allow oxygen (and other contaminants) to leech through the polycarbonate substrate layer.

In NIST tests, a CD totally submerged in clean water for 24 hours was found to be unreadable initially after removal and surface drying. It played normally, however, after 24 hours of drying out at approximately 70°F and 50% relative humidity (normal room conditions).⁴⁶

This surprising test result does indicate some of the resiliency of this storage media. However Byers indicates that if any contaminants or dissolved minerals are present in the liquid (not uncommon) they will be left behind after drying and will possibly react with the data, metal and adhesive layers.

As mentioned earlier, the organic dyes used in CD-R and DVD-R discs degrade over time regardless of optimum environmental conditions. The phase-changing film in RW and RAM discs degrade naturally as well and at a faster rate than the dyes in CD-R and DVD-R discs.

Sunlight has a profound negative effect on both writable and re-writable CD and DVD discs. The photons from direct sunlight are energetic enough to 'burn' the photosensitive dyes of CD-R and DVD-R discs, quickly making them unreadable (in a matter of days).

Heat build up as a result of direct or indirect sunlight (through the CD or DVD case) or other sources (radiator) has an effect on both photosensitive dyes and phase-changing films, degrading them quickly, making the disc unreadable. Extreme heat can also cause the disc to warp.

CD and DVD discs are very resilient to the effects of surface scratches. As the laser is not focussed on the surface of the disc but rather on the interior, scratches do not usually interfere with the process of reading or writing data. However on all types of CD discs the upper most layer (lacquer) which is usually covered by a label, is very thin. If this layer is scratched there will be either an immediate or a gradual effect on the metal layer directly below. The immediate effect can be damage to the metal layer causing permanent damage to the data in that area. Shallow scratches in the lacquer layer may in time expose the metal layer to the environment, accelerating its degradation.

⁴⁵ Byers, Fred R. "Care and Handling of CDs and DVDs" Page 9

⁴⁶ Byers, Fred R. "Care and Handling of CDs and DVDs" Page 18

3.4. Solid State Memory

Flash memory is a solid state storage device, which means that it contains no moving parts. This recent development has been implemented mainly for the use of digital cameras and video games consoles. Examples of Flash Memory are:

- CompactFlash, SmartMedia, and Memory Stick (all used in digital cameras).
- PCMCIA Type I and Type II memory cards - used in laptops.
- Any computer's BIOS chip.
- Video game console memory cards.

Flash memory works via a grid of transistors called control gates and floating gates, which store zeroes and ones by exchanging electrons, thus changing their relative electric charges. By measuring these relative negative or positive charges, data is accessed. Flash memory will retain data without any external source of power. The transistor states can be returned to an empty state by the application of an electric charge. Instead of erasing one byte at a time Flash Memory erases entire sections of memory (256 to 512 bytes), or even the entire chip, at once.

There has not been much testing of the long-term practical use of Flash Memory, especially the newer versions such as SmartMedia and CompactFlash cards. According to Howstuffworks.com Flash Memory cards are less rugged than other forms of removable storage and as such should be handled and stored with care.⁴⁷

3.5. Future Media

As has been shown in this chapter, storage media to date is, at its best and under optimum storage conditions, reliable for up to 75 years depending on the medium⁴⁸. At its worst it is completely unpredictable and easily degraded by environmental factors. Storage media formats developed in the near future will no doubt run the same gamut of predictability and would wisely be approached with caution until they have been rigorously tested in real conditions.

One of the next developments mooted for imminent release is holographic memory, which promises to condense vast amounts of information (1 terabyte or 1,000 gigabytes) into a sugar-cube-sized crystal.⁴⁹ This new technology will have to employ a light-sensitive ingredient to record information in 3 dimensions. Although this is an ingenious development, it was shown in the discussion on optical discs that light-sensitive media are extremely susceptible to the effects of

⁴⁷ Howstuffworks.com website, "How Flash Memory Works", <http://computer.howstuffworks.com/flash-memory.htm> accessed 24/02/04

⁴⁸ See Table 2: Sample Generic Figures for Lifetimes of Media, page 17.

⁴⁹ Howstuffworks.com website, "How Holographic Memory Will Work", <http://computer.howstuffworks.com/holographic-memory.htm>, accessed 24/02/2004

heat, and they degrade naturally over time, and so cannot be considered in any way a long-term solution.

The next aspect to be considered with regard to storage media is the hardware necessary to access the information stored. This will be discussed in Chapter 4 - Hardware Obsolescence.

4. Hardware Obsolescence

This chapter concerns the progress of hardware development and its impact on user's ability to access stored digital data. It is a progression from the issues of storage media degradation.

Data stored on specific media can only be accessed via the appropriate playback devices. For example, in order to access information stored on an 8-inch floppy disk, access to an 8-inch floppy disk drive is required. There are several reasons why this may not be possible.

- Worn-out hardware
- Technology discontinued by hardware producer

4.1. Worn-out Hardware

The drives necessary to access data stored on digital media wear out over time, just like any other machinery. Most computer hardware is not serviceable by the user, except through very limited cleaning tasks, such as running a drive head-cleaner in the hardware. Economically it has usually proven cheaper to replace worn out components than to attempt to have them serviced.

This is a feasible option until the manufacturer discontinues the product line, which appears to be inevitable for every storage product produced for the computer industry to date.

4.2. Technology Discontinued

Manufacturers of computer hardware components are in constant competition with each other to corner their share of the market. This has resulted in a healthy development of innovative storage solutions, each compressing more capacity into smaller and usually more affordable packages. At a certain point it becomes no longer economically viable to continue support or production of an older technology, and that technology then becomes obsolete.

The discontinuation of products may also happen through other market forces. The company may go out of business for a multitude of reasons. Market dominance by another company may force the company to shift emphasis to other product lines.

This may be the reason that the company Imation ceased producing the Imation SuperDisk™

120MB diskettes and drives as of May 2003⁵⁰, which were for a long time in direct competition with the lomega range of high capacity removable disk storage. It would appear that lomega has won that battle.

Even so, the lomega website contains a Legacy Storage page⁵¹ that lists diskettes and drives no longer produced or supported by the company. There are 14 separate headings covering 86 products altogether. This does not indicate that the storage diskettes are no longer accessible for all products listed here, but it does highlight the pace of their product development.

The example of the competition between lomega and Imation highlights a current storage media hardware obsolescence issue, and this should be considered in the broader history of computer hardware development. From the 1950s until today there has been a succession of hardware developments which has made the previous hardware 'obsolete' and lead shortly thereafter to a discontinuation of the 'obsolete' type.⁵²

Storage hardware does not always become obsolete because of the introduction of an intrinsically better technology. The computer marketplace is controlled by the same financial and perceptual rules as any other competitive market. Some technologies gain market dominance through early introduction and broad adoption and acceptance. For directly competing companies such as lomega and Imation, the technology may be equally advanced, but their market competition eventually drives one into a financially untenable position. Users of the Imation range of products are now looking hardware and storage media obsolescence in the face. According to the Imation website:

Please note, once Imation supplies of an obsolete product are depleted, we do not know which, if any, retailers might have products available. Therefore, Imation is unable to assist you in obtaining obsolete products.⁵³

4.3. Near Miss

Another large group of users had a recent close-up view of hardware obsolescence of a different kind. This involved the Apple MacIntosh (Mac). First released on the market in 1984, the Apple Mac was an instant hit with computer users due to its intuitive graphic user interface that hid the

⁵⁰ Imation website, "Imation SuperDisk Technology, Keeping You Informed"
http://www.imation.com/en_US/product.jhtml?Id=IM_FAM122 accessed 12/03/2004

⁵¹ lomega website, "lomega® Legacy Products : Discontinued Data Storage Drives from lomega", http://www.iomega.com/na/products/product_family.jsp accessed 12/03/2004

⁵² See Appendix # 6 - Computer storage timeline

⁵³ Imation website.

complexity of the DOS system common to most computers of that age.

Further releases followed building on strengths of the system. The Apple Mac was a single unit, produced in house, so the Operating System written for it was comparatively simple to implement and maintain. This is because the operating system did not have to be designed to recognise a multitude of varying hardware set-ups, as is the case with PCs, but rather had a predefined set of hardware components it was designed to recognise. As a result the Mac OS has a well-founded reputation for being a very stable system.

By 1997 however the business was floundering and many technical and financial experts were predicting the death of the company. The original founder Steve Jobs was hired back (he was fired in 1985) as temporary CEO in an attempt to inject new life into the company. His driving vision brought to the world the visually stunning iMac, and ushered in the current comeback that Apple Mac has been experiencing since.⁵⁴

Why did Apple Mac come so close to going out of business and leaving behind a vast tract of obsolete hardware, peripherals, software and data? Steve Jobs theorises that:

"The Mac-user interface was a 10-year monopoly," says Jobs. "Who ended up running the company? Sales guys. At the critical juncture in the late '80s, when they should have gone for market share, they went for profits. ... And then their monopoly ended with Windows 95. They behaved like a monopoly, and it came back to bite them, which always happens."⁵⁵

So this technology that was built on a very stable hardware and OS set-up did not automatically come to be the dominant and accepted computer technology.

Even though Apple Mac has now stepped back from the brink of extinction, it is another good example of the risks inherent in any dependence on a particular hardware system. Had they gone out of business, users the world over would eventually have lost access to replacement hardware, and shortly thereafter access to any digital data trapped in Apple Mac format.

⁵⁴ This history was from various sources, mainly:
MSNBC 2004 Newsweek article "OK, Mac, Make a wish" <http://msnbc.msn.com/id/4052227> accessed 12/03/2004

Flink, Chuck. "Falling on their Face: Six Incidents from Corporate History", May 27th, 2000
http://www.activewin.com/editorials/charles_flink/ink/23.shtml accessed 12/03/2004

Levy, Steven. "Hello Again", Newsweek 1998
<http://www.geocities.com/ResearchTriangle/2952/imacnews.html> accessed 12/03/2004

⁵⁵ MSNBC 2004 Newsweek article "OK, Mac, Make a wish" <http://msnbc.msn.com/id/4052227> accessed 12/03/2004

Mac floppy disks are formatted differently from Windows and as a result the Windows OS does not recognise by default files stored in Mac format. There are software applications available which enable Windows to recognise Mac files, but this does not guarantee the user will be able to open or manipulate those files as they may be dependant upon a Mac software package. This problem of accessing files of varying formats will be investigated in the next chapter, File Format Obsolescence.

5. File Format Obsolescence

This chapter will investigate the process and implications of file format obsolescence. File format obsolescence is a long-term storage issue that pertains mainly to proprietary file formats, as opposed to standard formats.

5.1. Official Standards and De-Facto Standards

A standards organisation is formed by a group of interested parties (usually companies producing the relevant product) in order to draw up accepted industry standards of development. With relation to the computer industry, various standard file formats and coding languages have been agreed upon which are then implemented and supported by many software packages. An example would be the Joint Photographer's Expert Group (JPEG) standard image format, the JPEG. The JPEG standard is recognised by the International Standards Organisation (ISO)⁵⁶. Another example is The World Wide Web Consortium (W3C)⁵⁷ which has recommended web standards, to which the browser software development companies are beginning to comply. These standards include specifications for HTML, XHTML and CSS.

De-facto standards come about as a result of general and widespread acceptance by the industry, not as a result of any standards planning process. This can come about as a result of a company releasing software which becomes very popular and therefore widely accepted by both users and the relevant industry. An example of a de-facto standard is the PostScript page description language (PDL) developed by Adobe Systems in 1982 as a language to print documents. Print companies using very high-resolution laser printers to produce camera-ready copy quickly accepted it.⁵⁸

Standard file formats (official or de-facto) have a greater chance of being accessible into the long-term future due to their broad acceptance and implementation by the computer industry. If a particular format, such as JPEG, is accessible by hundreds of different software packages, there is a greater likelihood that some of these applications (or their descendants) will be available to

⁵⁶ International Standards Organization (ISO) – a worldwide federation of national standards bodies from some 100 countries, one from each country. ISO's work results in international agreements which are published as International Standards. website <http://www.iso.ch/> accessed 11/03/2004

⁵⁷ World Wide Web Consortium (W3C) – develops specifications and guidelines to lead the Web to its full potential. As of 11/03/2004 the W3C has 371 members (companies). website <http://www.w3.org/> accessed 11/03/2004

⁵⁸ Webopedia.com website, <http://networking.webopedia.com/TERM/P/PostScript.html> accessed 11/03/2004

access the file format in the future.

5.2. Proprietary Formats

Proprietary file formats are privately owned and controlled. In computing terms this means that the file format often can only be accessed by the software that was used to create it. Since the format is not an open standard, other software developers have no way of implementing support for the format within their software applications without paying out royalty and other usage fees. If a proprietary format becomes so popular that it becomes a de-facto standard, then other software vendors will be forced to invest in making the format accessible by their own software.

The fate of any proprietary format is dependent on the fate of the controlling company. If the company ceases trading for whatever reason (bankruptcy, bought out, etc) then the software support for the file format will likely terminate shortly afterwards. If the controlling company decides to make changes to the format, the end users are directly affected by these decisions. This may lead to improvements that the end users find desirable, but it may also lead to increased software costs, forced upgrades from packages that the users are comfortable using, or other unpredictable results.

The Adobe Portable Document Format (PDF), although a de-facto standard for publishing and retention of documents, is still a proprietary format owned and controlled by a single company. The only way to access PDF documents is through the free (currently) Adobe Reader® software.

Another de-facto standard, proprietary file format is the Tagged Image File Format (TIFF), the copyright to which is also owned by Adobe Systems Inc. Many software packages developed by other companies can access TIFF documents, but at a cost that is under the control of Adobe Systems Inc.

A more obvious example of a proprietary file format is the Adobe Photoshop file. This contains information specific to the Photoshop software, such as layers and filters. As such this file format cannot be accessed by other software. A limited number of packages do have the capacity to import Photoshop files, for example Adobe AfterEffects can import Photoshop layers for use in video special effects. However this could not be considered authentic access as the file is not presented in its original format but rather is being reused for different purposes.

Proprietary file format obsolescence becomes an issue when stored digital data becomes inaccessible because the software necessary to interpret or decode the digital signal is no longer available. This may come as a result of several factors.

- Market Obsolescence - Software is removed from the market by the vendor.
- Operating System (OS) Obsolescence - Appropriate operating system is not available to run software.
- Hardware Obsolescence - Appropriate hardware is not available to run either the necessary operating system or the software.

5.3. Market Obsolescence

Market driven file format obsolescence usually happens as a result of the vendor releasing newer versions of any given software. For example, Microsoft devotes a page of their website to “Product Lifecycle Dates – Windows Product Family”⁵⁹. This page contains details of Windows products, the general release dates, and when support for the product was or will be retired. It also lists if an extended period of support has ended.

Market driven file format obsolescence may also be as a result of the vendor going out of business.

5.4. OS Obsolescence

The appropriate operating system is not available to run the software. Operating systems in general only support a certain level of backwards compatibility with older software packages. Likewise, newer software applications are designed to run on newer operating systems. To access proprietary file formats in the future, considerations may have to extend beyond just the necessary software package to include accessing the necessary OS.

The above mentioned “Microsoft Product Lifecycle Dates” webpage documents a minimum of 15 operating systems which Microsoft produced but no longer provides support for (as of March 11th 2004).

5.5. Hardware Obsolescence

Appropriate hardware is not available to run either the necessary operating system or the software. As strange as it may sound, many software packages, especially games, are hard coded to require specific hardware set-ups. Early software packages available on floppy disk are

⁵⁹ Microsoft website, “Microsoft Product Lifecycle Dates”, [http://support.microsoft.com/default.aspx?scid=fh;\[In\];LifeWin](http://support.microsoft.com/default.aspx?scid=fh;[In];LifeWin) accessed 11/03/2004

designed to be run from the floppy drive, usually designated as the A drive, and will not function properly when migrated to media that runs on a drive with a different designation.⁶⁰

It becomes obvious that file format obsolescence is only one part of a very large jigsaw puzzle (as was illustrated in Figure 1 – The chain of digital preservation). In order to access obsolete file formats, it is necessary to have access to the original software packages, which is also a further storage requirement. If these software packages are designed to run on specific Operating Systems (or narrow range of OS) then these also need to be archived. If they are tied to a specific hardware configuration, given that older OSs may not function on more modern hardware builds, then it becomes necessary to preserve and maintain the appropriate hardware, or else to develop emulators of the necessary systems. Emulators are software mock-ups of more primitive systems, and will be discussed in Chapter 7 - Digital Archiving Solutions.

Along with the software, OS and hardware specs, it will be necessary to have access to pertinent documentation for these systems, so that future archivists will be able to unravel the puzzle of putting them together in an operational fashion.

5.6. File Format Conversion

There is a possible short-term solution, which involves conversion of digital documents from an obsolete file format to currently supported file formats. This requires accessing archived files and then re-saving them using current software. For example opening a Photoshop version 1 document and saving it as a Photoshop version 2 document. During the next software cycle this is repeated from 2 to 3, 3 to 4, 4 to 5, etc. The shortcomings of this process quickly become apparent, as this is much more involved than the process of copying binary signals. Even when software vendors provide backwards compatibility for two or three versions (see 5.7 Backward Compatibility below), file format conversion still becomes an exponential workload as the number of documents in an archive grows.

Consider that Adobe Photoshop is still currently in use, but for how long will this continue? Many proprietary file formats are no longer in use because software companies come and go, and for the information contained in those files there may now be no way forward unless their format is supported by other software packages.

File format conversion can also lead to unpredictable results, as illustrated in the following quote:

⁶⁰ See Appendix # 3 - Farewell My Floppy

In one case involving FDA-mandated records of drug-testing, blood pressure numbers were randomly off by up to eight digits from those in original records following data transfer from UNIX platforms to Windows NT operating systems.⁶¹

5.7. Backward Compatibility

Backward compatibility refers to the ability of software packages or operating systems to access, run, recognise or otherwise process older file formats.

Backward compatibility in software packages can lead to code bloating. This is one reason why for example successive releases of the Windows operating system have required increasingly large amounts of hard drive space, as legacy code is included to allow the new system to access older file types. At some point the system performance will degrade to an extent which will demand a cut off point of backward compatibility support.

File formats change over time as a result of market driven competition amongst software developers and vendors. In 2004, many PC users create text documents using Microsoft Word, but during the 1980s the market leader was WordPerfect. Microsoft Word provides backward compatibility only to versions 5.x and up of WordPerfect. Prior to WordPerfect the standard application was WordStar, which is not supported by default by current text editing software packages. Emulators and conversion filters for WordStar can be downloaded and installed into packages such as Word, via the WordStar resource website⁶².

A similar situation exists for Microsoft Excel, whose users may have started with the original spreadsheet application VisiCalc (1979) before migrating to Lotus 1-2-3 (1983), and then Excel (Mac in 1985 and Windows in 1987) and QuatroPro (1989). QuatroPro never gained the popularity that Excel did, however the current version of Excel can open files created by both QuatroPro and Lotus 1-2-3 but not VisiCalc.

Thankfully for historians and archivists, interested individuals are making an effort to keep certain outdated software available on the Internet. So VisiCalc is available by permission of Lotus (who now own the copyright) via Dan Bricklin's website⁶³ (one of the creators of VisiCalc). VisiCalc is a DOS executable software package that can run on modern operating systems and weighs in at only 26.8 KB (very tight programming but it also has a minimal user interface).

⁶¹ Stepanek, Marcia, "Data Storage: From Digits to Dust", Business Week 20/04/1998, <http://www.businessweek.com/archives/1998/b3574124.arc.htm> accessed 20/02/2004

⁶² WordStar Resource website <http://www.wordstar.org/> accessed 11/03/2004

⁶³ Dan Bricklin's website <http://www.visicalc.org/history/vcexecutable.htm> accessed 11/03/2004

There is nothing inherently wrong with this progression of software packages, as each has added new features to extend users abilities to manage large volumes of information and to create professional publications. However, with each decision to upgrade to a newer software package comes the decision to either convert data to the new software format, or else to archive data along with the older package so as to allow access at some point in the future. As upgrades progress, say every 5 years, the amount of data requiring conversion grows unless a cut-off date is decided upon, in which case data before a certain period will effectively be lost as it will no longer be accessible.

5.8. Open Source Format Issues

Setting aside the issue of proprietary file formats for a moment, consider an open source, non-proprietary format. The World Wide Web's main display file format HTML is an open standard set by the World Wide Web Consortium (W3C). It is non-proprietary; it is a plain text (ASCII⁶⁴) file that should cross platforms and browser software transparently. Any company can write software to produce HTML files and any company can produce browsers to read HTML, and the file format will always stay the same, as a plain text file, and as such it should be protected from any form of software obsolescence.

However, browsers such as Netscape Navigator and Internet Explorer present HTML content differently and occasionally fail to present them at all due to the varying way that the browsers interpret the coded content. The 'browser wars'⁶⁵ encouraged a variety of responses from the website design community to this issue. Sloppy coding practice became acceptable due to the very loose interpretations of HTML by some browsers, HTML code which would then fail in other browsers. Professional website design firms implemented complex JavaScript coding targeting specific browser platforms in an attempt to serve up identical content to all browsers. This defeats the purpose of the universal file format that HTML was planned as.

Although the consistent visual presentation of textual content may seem a small point, the complete failure of some documents in varying browser versions means that it is not a straightforward process to access them when they are archived over long periods. Part of the archival solution for HTML documents may have to include some form of automatic HTML code 'tidy' software. This will have to be non-browser specific, since even though Microsoft's Internet

⁶⁴ ASCII - American Standard Code for Information Interchange: This is the de facto world-wide standard for the code numbers used by computers to represent all the upper and lower-case Latin letters, numbers, punctuation, etc. See Appendix # 1 - Definitions

⁶⁵ 'Browser wars' is a term commonly used to refer to the ongoing competition between Netscape Navigator and Microsoft's Internet Explorer. A very good article about 'browser wars' is available at Wikipedia, an online encyclopaedia. http://en.wikipedia.org/wiki/Browser_wars accessed 25/02/2004

Explorer may have market dominance in recent years, this may not continue in the future - nothing is certain.

Archivist would argue for exactly the opposite approach. The original HTML document should remain untouched and an attempt should be made to include information about what browser(s) the document was designed for, so that historians will be able to see the document as it was originally intended to be viewed, and as it was originally designed. Any code tidy software may implement unpredictable changes on the HTML documents.

These issues are beginning to be resolved with the implementation of XHTML and CSS, which attempt to separate content from presentation. The good thing about these new standards is that when properly implemented they will not fail in older browsers. There is less necessity for browser detection codes, which themselves can often cause pages to fail.

This is only one step towards avoiding file format obsolescence in a non-proprietary standard file format. It is in no way indicative of the general industry approach to file format support. As such, file format obsolescence will continue to be a major concern of any long-term storage solution.

6. Digital Archiving Issues

When considering the need to archive digital data, be it business records, cultural artefacts, or simply a child's early attempts at computer artwork, there are many factors to be considered beyond those of traditional archiving practice. These include digital storage media degradation, storage media and hardware obsolescence, and file format / software obsolescence (as outlined in other chapters). The concept of archiving is generally synonymous with concepts of preservation.

Preservation includes all the actions taken to extend the useful life of these materials and particularly the information contained on them. Therefore control, storage, handling and security are involved, as well as copying and reformatting or migration.⁶⁶

Digital archiving can also be seen as an act of preservation, but digital objects (be they images, text documents, databases, etc.) in an archive must also be accessible to be of any use. If a digital object is inaccessible it is for all purposes useless, as the binary signals that make up digital objects are not visible or meaningful to the naked eye.

Digital materials have been shown to require more immediate archival attention than their contemporary paper equivalents, due to their comparatively short media lifespan. As such guidelines and procedures need to be developed and applied much earlier after acquisition by an archiving institution. Such an institution may be a government-run archiving library, or it may be the creator of the digital materials.

The key requirements of any long-term digital archival solution will have to include the following components or considerations:

- Longevity / Life Expectancy
- Access / Inter-operability
- Authenticity
- Reference Structures and Naming Conventions
- Data Redundancy / Reliability
- Total Cost
- Ease of Migration, Refreshing and or Conversion

⁶⁶ Woodyard, Deborah. "Farewell My Floppy"

- Uses as Backup and Recovery
- Documentation

6.1. Longevity / Life Expectancy

Where and on what will the digital objects be stored? As has been shown, digital storage media are far from permanent with regard to long-term storage needs. Strict control of the storage environment can greatly extend the media lifespan, however due to hardware obsolescence issues, media migration will be an ongoing concern. For very large digital collections, the migration or refreshing of storage media may become a continuous process.

The lifetime of digital materials is dictated by many factors. These include degradation of the storage media itself (disks), obsolescence of media types and hardware, and file format obsolescence. Each of these factors must be anticipated and acted upon in a timely fashion or else the first indication of a problem may be the discovery that the material is inaccessible and useless. The goal is to be able to access digital data over very long periods of time, with no degradation of information content including formatting. As such the archive must always be aware of the current longevity status of any given digital documents in the collection.

6.2. Access / Inter-operability

A digital archive must ensure that the digital contents are protected from hardware and software obsolescence, so that the ability to access digital data content through successive generations of software and hardware development will be maintained. The solutions chosen to manage this situation must ensure long-term access while maintaining authenticity of the original digital materials.

6.3. Authenticity

Authenticity refers to several concepts when applied to digital materials. First the archive must ensure that the digital information is not edited or corrupted, purposely or unknowingly. This applies to the original document primarily, as any conversion involves change on a fundamental level. Check sums⁶⁷ can be used to ensure that the digital data has been preserved intact and without errors. This must be an ongoing process as corruption of the binary signal can happen unpredictably, for example due to storage media degradation.

Authenticity also refers to the original presentation of the digital information, be that the formatting

⁶⁷ Checksum: A computed value which depends on the contents of a block of data and which is transmitted or stored along with the data in order to detect corruption of the data. Definition by Hyperdictionary.com, accessed 08/03/2004 See Appendix # 1 - Definitions

of a text document, the logical layout and interpretation of a database of information, or the interactivity of dynamic material. Losses of text formatting, even though the full content of text remains, can lead to profound losses of information. This is because text documents often contain information contained solely in visual cues, such as indenting, highlights, italics, etc.

Figure 3 and Figure 4 below illustrate an example of the loss of authenticity.

I	II	IIIb	IVb	Vb	VIb	VIIb	VIIIb	Ib	IIb	III	IV	V	VI	VII	0		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac**	Rf	Db	Sg	Bh	Hs	Mt	110	111	112	113					
Lanthanides *				Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Actinides **				Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Figure 3 - Periodic table created with a non-spatial font.⁶⁸

I	II	IIIb	IVb	Vb	VIb	VIIb	VIIIb	Ib	IIb	III	IV	V	VI	VII	0		
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	La*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Ac**	Rf	Db	Sg	Bh	Hs	Mt	110	111	112	113					
Lanthanides *				Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Actinides **				Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Figure 4 - The same periodic table rendered with a spatial font.⁶⁹

Authenticity also applies to interactive digital documents. Interactivity can take many forms when applied to digital documents, and most often involves dependency on other documents, software or even hardware set-ups. Games can require specific minimum hardware specifications and software plugins. HTML documents require that relative linking structure be maintained, i.e. if

⁶⁸ Dr Raymond J van Diessen and Dr. Johan F Steenbakkers, "The Long-term Preservation Study of the DNEP project - an overview of the results", IBM Netherlands, Amsterdam, December 2002

⁶⁹ Dr Raymond J van Diessen and Dr. Johan F Steenbakkers, "The Long-term Preservation Study of the DNEP project"

target documents are renamed then the dynamic links will no longer be valid. There are numerous codecs⁷⁰ associated with video and audio files, and those files will require players that recognise these codecs to decode their formats.

Another consideration applied to authenticity surrounds issues of security and access. The archival system must safeguard against malicious modifications of digital documents. The system must be able to recognise and recover altered documents. Safeguards must be considered, such as how to restrict access to the master copies of digital data without restricting their usefulness.

Digital information stored over a long period of time may undergo many transformations in the archiving process. These involve migration from deteriorating storage media, refreshing to new media, and conversions from obsolete file formats. To maintain authenticity of the digital document, these changes over time require documentation. Ideally the original digital data should also be preserved (as opposed to converted), in the hope that obsolete file formats may become accessible in the future using technologies such as emulation to recreate the original software and hardware environment.

6.4. Reference Structures and Naming Conventions

How will the digital objects be referenced for future users to find? This will require strict information management, and must also maintain authenticity (see above). Outside of a library or archive setting, the person creating the digital document chooses the digital file names and it cannot be assumed that they have observed any strict naming convention. A digital file name may or may not contain useful information regarding the object itself. If new naming conventions are applied to the digital file, how will authenticity be maintained?

Consider that on a book shelf publications may be arranged in alphabetical order according to title or author. The digital environment allows multiple ways of arranging and searching for any given collection of documents. The referencing and naming conventions can be contained in a separate document, a catalogue, which then points to the archived digital files. There will be a need to keep this catalogue document current as digital documents are migrated to newer storage media.

6.5. Data Redundancy / Reliability

The nature of digital data allows for any number of identical copies of the data to be produced.

⁷⁰ CODEC - Acronym for coder-decoder. An electronic device, circuit, or software that converts digital signals to and from analog. Usually also includes digital compression technology for added efficiency. Different codecs may provide different efficiency, quality, and features. Definition by Hyperdictionary.com, accessed 08/03/2004

Therefore the archivist is not solely dependent upon a single object stored at a specific geographic location, and can implement a strategy of redundancy to ensure long-term preservation of digital information. Master copies can be held in controlled long-term storage environments, and users can access circulation copies for which long-term use is not an issue. If the data gets destroyed or corrupted in the circulation copies, new copies can be made from the masters.

Further redundancy can be implemented by ensuring that multiples of the master copy are stored at disparate geographic locations, to protect digital information from destruction due to single events, such as natural or man-made disasters (political terrorism and other malicious acts).

6.6. Total Cost

The archival solution must be of a reasonably affordable cost over the long-term. Expenses to be considered include equipment, maintenance, storage, migration, and conversion. Every aspect outlined above is accompanied by real world expenses, such as initial outlay for the archiving system, and ongoing technical upgrades. If outside vendors are employed to manage specific aspects, then these will become long-term expense commitments.

Will the archived information be able to survive economic fluctuations outside the control of the archive itself? If the digital information is dependent upon a third party vendor at some stage in the management process, there is the risk that such a company may go out of business. Is the archiving system stable enough to withstand long periods of dormancy caused by lack of funding? This last point is worth considering, as many archaeological finds of great cultural value have lain dormant for centuries or millennia.

6.7. Ease of Migration, Refreshing and or Conversion

The level of automation implemented in handling a digital archive's migration, refreshing and conversion needs will directly impact on its efficiency. To recap and avoid confusion, migration refers to the transfer of digital signals from one storage media to another, for example tape to CD. Refreshing involves transferring digital signals to an identical newer storage media, i.e. tape to tape, CD to CD, etc. Conversion involves changing the digital file format to another, such as converting from a Word file to simple text file or a PDF document. Conversion almost always involves the loss or corruption of information (see 6.3 Authenticity above).

All acquired digital documents will need to be tested against the archival system and rated for ease of implementation. That is to say, does the archive recognise the digital file format and storage media? How difficult will it be for future users to access this specific file (if proprietary)? Is

the archive in a position to convert or migrate the digital data without corrupting the original information? How difficult will it be to implement this conversion or migration? How often will the archive need to refresh this digital data? These factors will have a large impact on the success of an archival system.

6.8. Uses as Backup and Recovery

Along with the long-term preservation and access to digital documents of cultural or historic significance, digital archives may be implemented to provide backup services for organisations' daily digital information functions. In the event of a localised disaster, will companies be able to access recent (daily or weekly) full copies of important data and digital services efficiently?

6.9. Documentation

Throughout the lifespan of the preserved digital data, documentation on the various integral technologies involved with its production must be preserved. The term 'production' here refers not only to the initial creation and dissemination of the digital data, but also to the many various stages of preservation it must be put through. This requires documentation on any migration, refreshing or conversion processes and technologies, software and hardware documentation, as well as encoding documentation where applicable. This is especially true for any form of proprietary file format and custom made software or hardware – as is often the case with research institutions. The nature of digital data dictates that every digital file is simply a meaningless string of zeroes and ones until it is run through the appropriate decoding software.

Documentation on hardware is essential when it comes to recreating or emulating hardware systems, as many digital applications are hard coded to require specific system set-ups.

All of the major digital archive requirements outlined in this chapter will have to be effectively addressed when planning for a long-term solution. Digital technology is a very unique medium of information storage and communication, with its own unique requirements for long-term preservation which go beyond those of a static object which simply requires preservation of the physical medium. Digital technology requires strictly accurate preservation of the message, which is the information contained within the medium.

7. Digital Archiving Solutions

This chapter investigates some of the commercial and research solutions or recommendations to-date for the long-term storage of, and access to, digital information. None of the solutions introduced here are complete in and of themselves, as they do not fulfil all of the archive issues introduced in the previous chapter. However, they each have aspects of merit, and could present a single solution through a combination of their methods.

7.1. Hybrid Solutions

IT Storage Online, a website that publishes articles of interest to IT storage professionals, defines a combination of solutions as a 'Hybrid Solution'⁷¹. They recommend hybrid imaging (the transfer of digital documents to microfilm or other analog output) as a type of compromise. According to their definition it allows organisations the benefits of quick access to digital documents and a long-term storage solution that is *technology independent* (my italics).

Hybrid imaging incorporates digital and analog document records (microfilm) as part of the records-retention strategy. Hybrid imaging can digitally capture data for electronic output or microfilm archiving. ... Microfilm provides technology-independent, reliable storage copies for the less active stages of the document lifecycle, which may span years, decades or - in the case of permanent records - centuries.⁷²

7.2. Digital - Analog - Digital

The Eastman Kodak Company outlines its hybrid solution to the need for long-term retention of digital information in its document "Digital Insurance for Information at Risk - A strategic overview of digital preservation", published in 2000. Their approach is described as "Digital - Analog - Digital, Kodak's two-way street to digital preservation"⁷³. Briefly, they recommend outputting all digital documents into analog form (printed either on paper or microfilm) for long-term storage at a separate facility. When needed in the future, these 'images' can be re-digitised (scanned) for use as digital documents. They also indicate that image capture from original paper documents can

⁷¹ "Protecting Knowledge Assets", IT Storage Online, published 1/31/2003, accessed 12/08/2003. <http://www.itstorageonline.com/contents/news/>

⁷² "Protecting Knowledge Assets", IT Storage Online

⁷³ Lawrence, H. Andrew. "Digital Insurance For Information At Risk. A strategic overview of digital preservation", Eastman Kodak Company, 2000

also be implemented into the process. See their flowchart below.

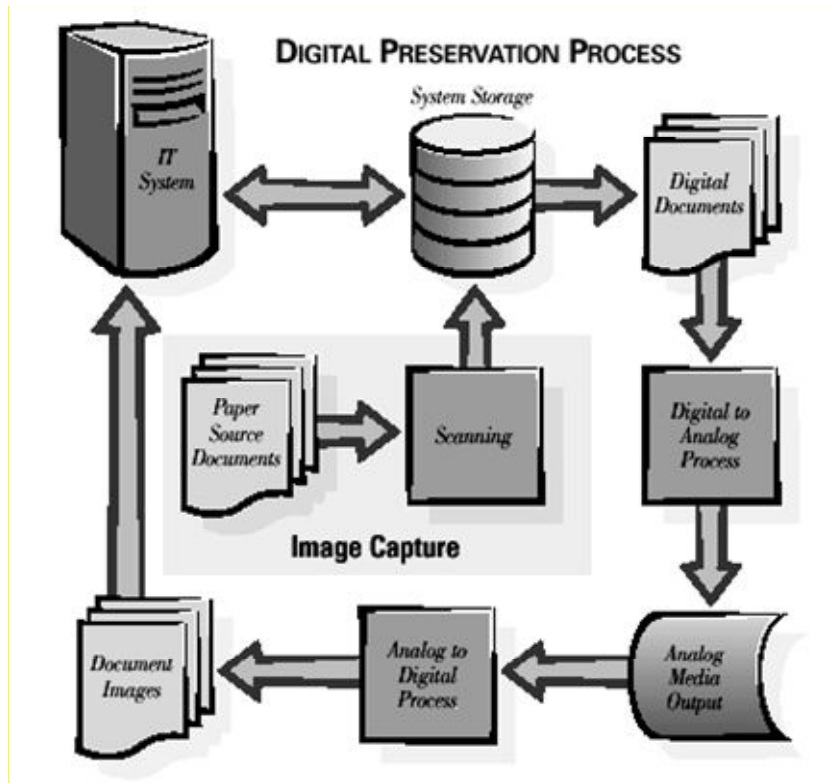


Figure 5 - Eastman Kodak Company's 'Digital-Analog-Digital' preservation process.⁷⁴

7.3. Digital To Analog Micro-imaging

Norsam Technologies produce digital-to-analog output in the form of their patented HD-Rosetta archival preservation process. This process records microfilm-like images onto a metal disk, composed of a thin layer of nickel or stainless steel on a silicon substrate. These discs range in size from 4" x 6" and recording 540 images, to 2" square or round and recording up to 200,000 images per disc. Norsam Technologies use photolithographic processes for the less dense discs, and focused ion beam techniques for the discs containing high densities of images.

Tests run by the Los Alamos National Laboratory in 1999 confirmed Norsam Technologies' claims that the discs easily survive temperatures up to 300°C and are highly resistant to the corrosive effects of saltwater and other forms of humidity. One test result tentatively showed that the discs could survive up to 400 years in a salt-water environment, but this was dismissed as extrapolation of results over very long periods is not accurate due to high probability of local chemistry changes over time.

⁷⁴ Lawrence, H. Andrew. "Digital Insurance For Information At Risk"

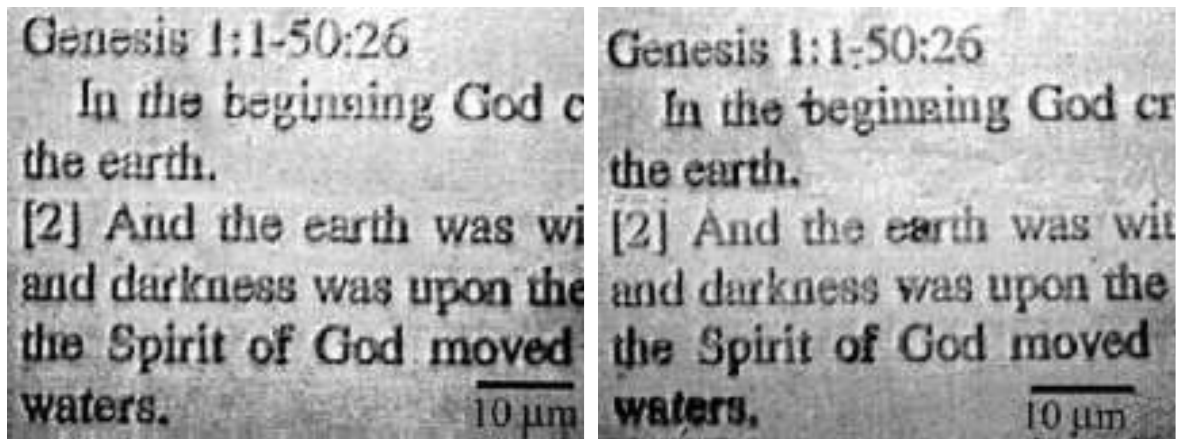


Figure 6 - The image left is the original text seen at 4,500 times magnification. The image right is the same text after exposure to 300°C (570°F) air for 24 hours. Each disc is 2.2 inches in diameter and contains approximately 9,000 pages of text or images.

Norsam Technologies claim that the HD-Rosetta metal discs are immune to technology obsolescence, but do acknowledge that higher density discs will require sophisticated microscopes for reading purposes. They claim that lower density discs can be read with “a lens”⁷⁵.

These companies, Eastman Kodak Company, IT Storage Online, and Norsam Technologies, are promoting a digital to analog solution which is only suitable for static documents. These solutions may be appropriate for documents that contain only static text and static images, but they will not be appropriate for digital documents that contain hyper-links, animations, video and audio components, dynamic data that responds to user input, games and other software. When the documents are output to analog form, they effectively become images and lose any inherent information regarding their content. They will not provide the scope of search facilities available to digital documents, whereby a collection of documents can be quickly searched for a word or string of characters. They lose any dynamic functionality they may have possessed, such as hyper-links. As an image, the information hidden behind a hyper-link will be lost unless this issue (and many similar ones) is addressed.

7.4. RAID Systems

Another storage solution proposed by many computer hardware manufacturing companies is the RAID System, which is an acronym for a Redundant Array of Independent (or Inexpensive) Drives. RAID is an assembly of disk drives, a disk array that operates as a single storage unit.

Spreading data across an array of disks does not increase reliability; if anything it drastically reduces reliability, since the probability that any one drive will fail grows. If any single disk fails then access to all the data stored across the array is lost. In an array of disks, the possibility of

⁷⁵ Norsam Technologies website, <http://www.norsam.com/rosetta.html>, accessed 20/02/2004

any disk failing raises exponentially as new disks are added, given that they are all the same type of disk and have the same mean time to failure (MTTF). MTTF is an estimate given by the manufacturer of the minimum guaranteed working lifespan of the drive. The original calculations by Patterson, Gibson and Katz in their paper, "A Case for Redundant Arrays of Inexpensive Disks (RAID)"⁷⁶ show that:

$$\text{MTTF of Disk Array} = \frac{\text{MTTF of a Single Disk}}{\text{Number of Disks in the Array}}$$

For example if one disk has a MTTF of 30,000 hours (more than 3 years), then an array of 2 disks would have an average MTTF of 30,000 hours divided by 2 or 15,000 hours. Extrapolating further, a 10-disk array MTTF is 300 hours (less than 2 weeks) and 100-disk array MTTF should only be 30 hours!⁷⁷ These calculations may appear to fly in the face of logic, but they are based on statistical probability.

The strength of RAID systems lies in redundancy, and differing configurations of RAID are classified as 'levels'. These configurations make use of extra disks containing redundant information so as to recover information if (and inevitably when) a disk in the array fails.⁷⁸

- RAID 0 uses striping⁷⁹, and does not provide redundancy. If one drive fails, the entire array will fail. RAID 0 is used to boost performance.
- RAID 1 uses mirroring⁸⁰ and provides 100% redundancy. It also however requires twice the capacity, i.e. one disk mirroring each disk in the array, so the cost doubles for any given storage array.
- RAID levels 3, 4, 5, 6, 7 and 53 contain redundant data in the form of parity⁸¹. Depending on the level (array configuration), parity information is written to other drives (either a dedicated drive(s) or across the entire array) so that when a drive fails, the data lost can be recovered from the parity information.
- RAID levels 10 and 0+1 combine mirroring and striping, without parity.

The different RAID levels provide varying degrees of redundancy protection and system performance, so that RAID level use depends upon the requirements of the situation, i.e. video

⁷⁶ David A. Patterson, Garth A. Gibson and Randy H. Katz. "A Case for Redundant Arrays of Inexpensive Disks (RAID)" page 110. SIGMOD Conference 1988

⁷⁷ David A. Patterson, Garth A. Gibson and Randy H. Katz. "A Case for Redundant Arrays of Inexpensive Disks (RAID)"

⁷⁸ See Appendix # 5 - Definitions of RAID levels

⁷⁹ Striping – A process whereby segments of data can be written to multiple physical devices in a round-robin fashion. Definition adapted from <http://www.hyperdictionary.com> 09/03/2004

⁸⁰ Mirroring - Writing duplicate data to more than one device, in order to protect against loss of data in the event of device failure. Definition adapted from <http://www.hyperdictionary.com> 09/03/2004

⁸¹ Parity - an error detection procedure in which a bit (0 or 1) is added to each group of bits so that it will have either an odd number of 1's or an even number of 1's; e.g., if the parity is odd then any group of bits that arrives with an even number of 1's must contain an error Definition adapted from <http://www.hyperdictionary.com> 09/03/2004

post-production versus e-mail servers. RAID systems are further enhanced when combined with hot-swappable disks. These standby disks are engaged when a drive failure is detected, automatically recovering the lost data and allowing the entire system to continue uninterrupted.

RAID is not being suggested as a long-term storage solution, as it is still a form of magnetic media and susceptible to the same environmental and inherent weaknesses as was outlined in the chapter on digital storage media degradation. RAID systems are implemented for their processing power and reliability, not as long-term (static) storage.

However it is worth making note of due to its capacity for Extended Data Availability & Protection (EDAP). EDAP is “the ability of a disk system to provide timely, continuous, on-line access to reliable data under certain specified abnormal conditions.”⁸² These abnormal conditions include⁸³:

- Internal failures of the disk array.
- External failures of the equipment attached to the array.
- Environmentally caused failures, such as floods, earthquakes, fires, sabotage, terrorism, etc.
- Replacement periods, the intervals during which failed equipment is being replaced and the system is vulnerable.
- Vulnerable Periods, during which the system has invoked all its possible error correction mechanisms and has no further fallback position. The system is vulnerable to further failures and operates at less than optimum performance.

That any system could be expected to provide “timely, continuous, on-line access to reliable data” during the aforementioned abnormal conditions demonstrates extreme resiliency, a condition which should be high-up in the list of considerations for any long-term archival solution.

The RAID Advisory Board⁸⁴ (RAB) introduced improved classifications of RAID systems in 1996 to include the EDAP specifications. RAID systems are classified according to three headings:

- Failure-resistant systems that protect against data loss.
- Failure-tolerant systems that protect against loss of data access.
- Disaster-tolerant systems that are separated into independent zones which can provide continuous data access.

As of May 26th 2002⁸⁵, the RAB website listed 6 companies with a total of 11 RAID systems

⁸² USByte.com website, “RAID Systems” http://www.usbyte.com/common/raid_systems.htm accessed 09/03/2004

⁸³ USByte.com website, “RAID Systems”

⁸⁴ RAID Advisory Board, <http://www.raid-advisory.com>. The RAB is a group of companies chartered to developing common definitions for the RAID levels. This organisation appears (from other sources) to have existed from 1992 until May 26th 2002, their last website entry at The Internet Archive Wayback Machine (<http://www.archive.org/>). Their website is now unavailable (domain for sale) and further searches have been fruitless.

⁸⁵ The Internet Archive Wayback Machine <http://www.archive.org/> accessed 09/04/2004

classified as Disaster-tolerant. These characteristics are displayed by another widely used system, the World Wide Web, no doubt partially through the implementation of RAID technology. The Internet stands as a prime example of new media technology that has proven very robust on two occasions of extreme conditions, the September 11th attacks on New York⁸⁶ and the August 2003 rolling blackouts in the north-eastern US and south-eastern Canada⁸⁷. However it could also be correctly argued that this robust functionality is a result of the Internet's overall implementation and not a result of any individual server hardware or software design. The robust nature of the Internet can be pinned on the communications protocols used to send digital signals from point to point, not on any hardware, storage medium or software.

The resiliency of the World Wide Web to continue to respond to information requests during major disasters involving local nodes (e.g. New York) suggests several further technologies to consider, Storage Area Networks (SAN) and Network Attached Storage (NAS).

7.5. SAN & NAS

Network Attached Storage (NAS) is composed of disk drives located on a network through which information is accessed and shared via a dedicated file server for clients connected to a LAN / WAN⁸⁸ or the Internet. A Storage Area Network (SAN) is a network of storage devices through which volumes of information are accessed via application servers. The distinction between SAN and NAS is often confused, understandably, and now these network storage systems are converging as businesses employ both methods for their increasing information management needs.

The main advantage of SAN is that storage devices can be directly connected as a network, rather than requiring an intermediary server for communication. Data can move freely and directly between storage devices, disk to disk, disk to tape, and tape to disk. This is often referred to as serverless-backup.⁸⁹ A server can initiate data transfer (backup and restore) procedures, but this transfer does not affect the server as the storage devices are on a separate network loop. As a result there is no live network 'downtime' for the process.

The Irish government has committed to putting in place disaster recovery plans for the storage of

⁸⁶ "The Internet Under Crisis Conditions: Learning from September 11", National Academy Press www.nap.edu

⁸⁷ James H. Cowie, Andy T. Ogielski, BJ Premore, Eric A. Smith and Todd Underwood, "Impact of the 2003 Blackouts on Internet Communications", Preliminary Report, Renesys Corporation, November 21 2003

⁸⁸ LAN – Local Area Network restricted geography usually to less than 1km radius.
WAN – Wide Area Network, geographically greater than 1km radius. Definitions adapted from <http://www.hyperdictionary.com> 09/03/2004

⁸⁹ Booyesen, Vic. "Information and the Archiving thereof", page 20

vital government data and services based on SAN. All government departments are required to have their data duplicated to at least two separate locations, as well as mirroring vital applications necessary to the department services. Up until recently this involved the nightly delivery of backup tapes to a remote secure site. However a major new infrastructure that is supporting this initiative is the fibre network connection to almost all public-sector agencies. This allows departments to transmit high bandwidth copies of sensitive data securely to alternative sites, and also allows them to re-route vital service applications in the event of local difficulties.

For example, the Department of Finance now uses a SAN supplied and supported by Hewlett-Packard, with a three-terabyte capacity server array located at the Merrion Street headquarters. This primary SAN is linked to a secondary SAN, which duplicates all data to a remote disaster recovery site.⁹⁰

As with the discussion of RAID, the SAN itself is not being implemented for long-term storage of information, but rather as an administrative resource. The SAN system is worth making note of because of the transparency of the system, another one of the main archival requirements (ease of use and automation). The SAN system also allows for the introduction of redundancy and reliability, notions which are necessary for long-term storage of and access to information.

7.6. Virtualisation

The next logical progressions from SAN and NAS are two very similar but extended concepts, virtualisation and grid computing. In the same way that a RAID system is presented as a single storage unit even though it is composed of an array of disks, virtualisation software presents a SAN to a server as a single 'pool' of storage resources. Virtualisation software hides the complexity of separate physical locations and disparate formats of a distributed system from the server, presenting it instead with a single interface to administer the collective SAN resources. This allows the server to manage the available storage, regardless of physical set-up, allocating capacity to applications as necessary. This effectively turns a SAN into a single, very powerful, processing and storage unit.⁹¹

7.7. Grid Computing

Grid computing uses virtualisation to leverage multiple SAN and NAS systems effectively for use by multiple users. Grid computing and virtualisation are intertwined concepts, in that one depends

⁹⁰ Faughan, Leslie. "Public sector is ready for disaster", Digital Ireland, The Irish Independent Newspaper, February 2004.

⁹¹ Bradbury, Danny. "Virtually the next big thing", Computer Weekly, October 10 2002, http://www.findarticles.com/cf_0/m0COW/2002/_Oct_10/92671444/print.jhtml accessed 10/03/2004

on the other for its successful implementation. Whereas virtualisation is implemented as a management tool for a SAN or NAS system, grid computing uses virtualisation to manage many distributed network systems as a single entity.⁹²

Grid computing allows multiple users to access and share data. It gives users transparent access to vast number crunching potential for data processing and complex system modelling, as well as seamless (authorised) access to any application running anywhere on the grid. Theoretically and ideally, grid computing would not just include networked storage devices, but any type of digital device connected to the network. These could be workstations, desktop PCs, data vaults, supercomputers, mainframes, remote sensors and other devices.

The concept of grid computing is a natural development of other networked processes already in use today.⁹³

- Distributed computing - a network system where multiple computers process data as a virtual single unit to formulate a result.
- Metacomputing – networking supercomputer centres over high-speed connections.
- Cluster computing – networking desktop PCs together to emulate mainframe or supercomputer processing power.
- Peer to peer computing – a process whereby individual computers can directly connect with other computers via a network such as the Internet and share files. Napster is a good example of peer to peer computing
- Internet computing – the best example is SETI@home⁹⁴. This screen-saver software is installed on individual PCs. The software scavenges idle time in order to process information for the SETI@home project. The software only runs when the computer is idle, at which point it initiates an Internet connection, downloads a small chunk of data to process and uploads the results. There are currently at least 10 such projects running on the Internet.⁹⁵

7.8. Global Grid Computing

Data Grids and Global Grids are extensions of grid computing, effectively using the same technology under a different name to define what the aims of the group involved are. This is because there is no single 'Grid' as of yet; instead there are separate grids, private, governmental, regional and global in structure. This area of computing is being examined and

⁹² IBM website, "What is grid computing", http://www-1.ibm.com/grid/about_grid/what_is.shtml accessed 03/03/2004

⁹³ Grid Café website, "What is the grid?", <http://gridcafe.web.cern.ch/gridcafe/whatisgrid/reality.html> accessed 15/03/2004

⁹⁴ SETI@home 1,844,370.5 years worth of CPU processing time has been accessed by the SETI@home software running on 4,912,815 computers worldwide since July 1999, as of 10/03/2004. <http://setiathome.ssl.berkeley.edu/> accessed 10/03/2004

⁹⁵ Grid Café website, "Grid projects in the world", <http://gridcafe.web.cern.ch/gridcafe/gridprojects/athome.html> The areas of research are: Climate prediction, Cancer (2), Protein folding (2), Genome mapping, AIDS drug testing, Mersenne prime numbers (mathematics), smallpox drug testing, and SETI (search for extraterrestrial life).

implemented extensively by research communities and corporations all over the globe. The Global Grid Forum was formed as a result of conversations and workshops held from 1998 to 2000 concerned with various aspects of grid computing.

The Global Grid Forum (GGF) is a community-initiated forum of thousands of individuals from industry and research leading the global standardization effort for grid computing. GGF's primary objectives are to promote and support the development, deployment, and implementation of Grid technologies and applications via the creation and documentation of "best practices".⁹⁶

The goal for many grid computing users and researchers is to implement a single global grid, consisting of every networked digital device connected together. This sprawling conglomerate of digital devices would have unimaginable processing power and be open to all, with appropriate security mechanism implemented of course.

7.9. How does this tie in with Digital Archiving?

As an extension of SAN and NAS data storage systems, a global grid with a truly transparent system of access and storage (virtualisation) could supply the individual user as well as large institutions with an extremely robust storage environment. This storage environment by its nature would automatically refresh over time, backing up information to new systems as hardware comes online, hot-swapping to new storage devices as older devices fail, in a RAID type process.

The global collection of computing devices would become one immense hot-swappable pool of storage and processing power. No one need ever know exactly where in geographic space specific data resides; just that it is being administered, mirrored, and is continually available.

Grid computing as it stands today is not being touted as a long-term storage solution, but rather an environment where colleagues can share and process data communally. Long-term storage by most institutions and commentators is viewed as 'static', not live. However grid computing should be considered as part of the hybrid solution, as it has many qualities which are in line with those of a long-term information archive.

Global grid computing won't come about tomorrow or the next day, as it is dependent upon further technological developments and infrastructures. Global grid computing will require:

- Affordable (or free), high bandwidth, always on, guaranteed access for all.

⁹⁶ Global Grid Forum website, <http://www.gridforum.org/> accessed 10/03/2004

- Unparalleled advances in networking security, to guard against malicious interference.
- Unparalleled advances in virtualisation capabilities to administer an immense and dynamically changing storage space.
- Unreserved and guaranteed backing by governments and institutions, as they will be no doubt supplying the bulk of the online processing power and storage resources in the form of hardware and infrastructures.
- Users accessing the grid would be expected to invest in and commit certain minimum specifications of hardware to the grid. That is to say, in order to avail of the grid a user must contribute to the grid.

All of this sounds very altruistic, and it would have to be for the system to work. However, the immense computing possibilities would be a major incentive for companies and research institutions to get involved on the ground level. The issues of security will dictate whether such a system ever gets off the ground. There are many self contained 'grids' in operation already⁹⁷, but they succeed because there is no outside (unauthorised) access available, and so the possibility of malicious damage is decreased.

For the near future then, global grid computing is not an immediate solution, merely a component to keep in mind for future use.

Several archival issues still need a component for the overall hybrid solution. These are authenticity and access. These two concepts outlined in the chapter on archival issues are inter-related. The digital data must remain in its original form in order to preserve its authentic functionality, presentation, etc. as an unchanged (maliciously or accidentally) object – i.e. the checksums must add up. The digital information must also remain accessible into the future. The digital documents have to be preserved in such a way that future computer systems will not be blocked by software and hardware obsolescence factors from accessing those documents.

7.10. Authenticity & Access through Emulation

IBM worked with the National Library of the Netherlands (Koninklijke Bibliotheek, KB) from 2000 to 2002 to develop a "Digital Information Archiving System" (DIAS)⁹⁸. The IBM DIAS solution focuses strongly on the conceptual models necessary to a long-term solution. The IBM DIAS is not a complete system, in that they have delivered a working system to the KB but have proposed developing further sub-systems to be integrated in the future. The IBM DIAS is basically a management software solution, but the further sub-systems proposed introduce radical ideas concerning authenticity and access.

⁹⁷ Global Grid Forum website

⁹⁸ Dr Raymond J van Diessen and Dr. Johan F Steenbakkens, "The Long-term Preservation Study of the DNEP project"

The IBM DIAS maintains authenticity by preserving the original digital data. This may sound straight forward, but what it implies is that files are never converted to newer formats, which would impact on their authenticity as shown in the chapter on archival issues. If the file formats are never converted, then how is continued access beyond software and hardware obsolescence maintained? Access is maintained through a type of emulation proposed by IBM.

A unit of storage within the DIAS is defined as an Archival Information Package (AIP). The AIP has an XML (ASCII text) table of contents containing information on the digital objects contained within the AIP. This XML document contains information about the name and file type of all the digital objects in the AIP. The file type is a number that specifies the view path (see below) associated with a specific file.

The view path is a collection of technical metadata regarding the bit stream, the viewing application (such as WordPerfect), and the necessary OS and hardware platform. The view path is described as a series of discrete layers necessary to access any given file. This technical metadata is stored as a separate AIP, and new AIPs are created as necessary. The original digital document is never changed, just the XML table of contents, the view paths, and the technical metadata. The view path is used in the future to create the emulations necessary to access the digital data. By separating the technical metadata from the original digital documents, the technical metadata can be updated without risking possible corruption of the original digital documents.

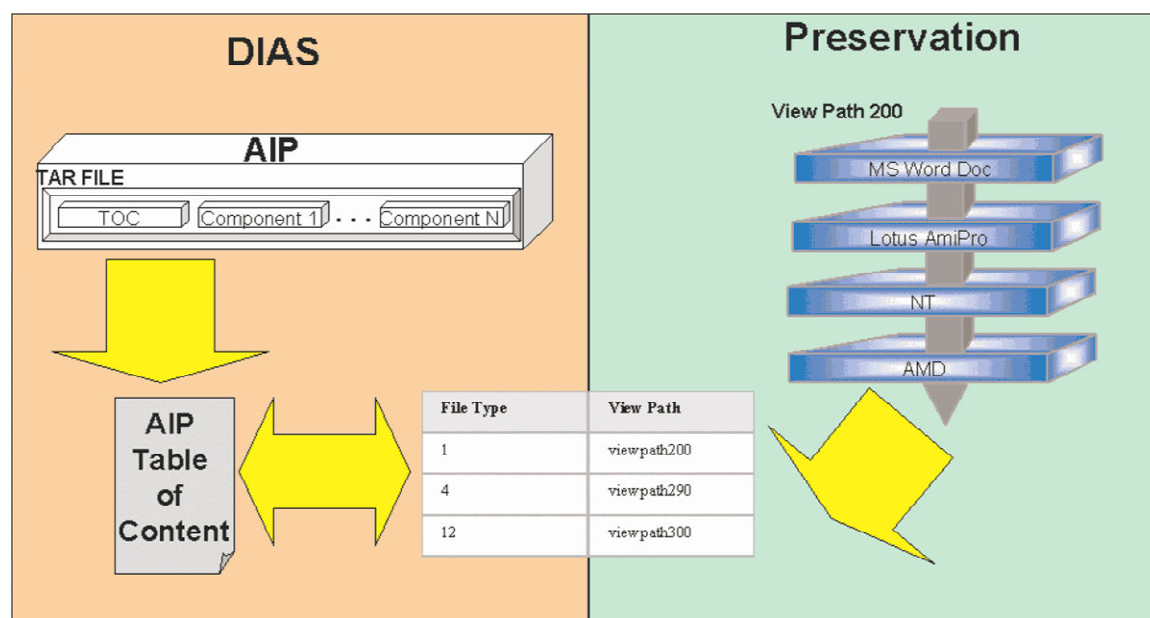


Figure 7 – Overview of the IBM KB DIAS preservation flow path system⁹⁹

⁹⁹ Dr Raymond J van Diessen and Dr. Johan F Steenbakkers, “The Long-term Preservation Study of the DNEP project” Page 25

The IBM DIAS combines this AIP file path system with a proposed emulation package called a Universal Virtual Computer (UVC). The UVC is built out of a simplified, standardised architecture, for which hardware emulators are written, preferably of hardware specifications before they become obsolete. These emulations are designed for the UVC instead of for a particular physically existing computer. In this way, in the distant future an emulator can be built to run the UVC itself, upon which all the pre-made hardware emulators can function. According to IBM:

Because the UVC instruction set is so simple, it is relatively straightforward to write a UVC emulator for any given computer.¹⁰⁰

By emulating hardware specifications, there is no need to build software emulators. The original software, such as operating systems and applications, can be installed onto the hardware emulation. To access any particular file then the future user will require:

- An emulator to run the UVC
- The UVC
- The original digital document with its associated view path and technical metadata
- The appropriate hardware specification emulation and the original software and OS necessary to access the digital document.

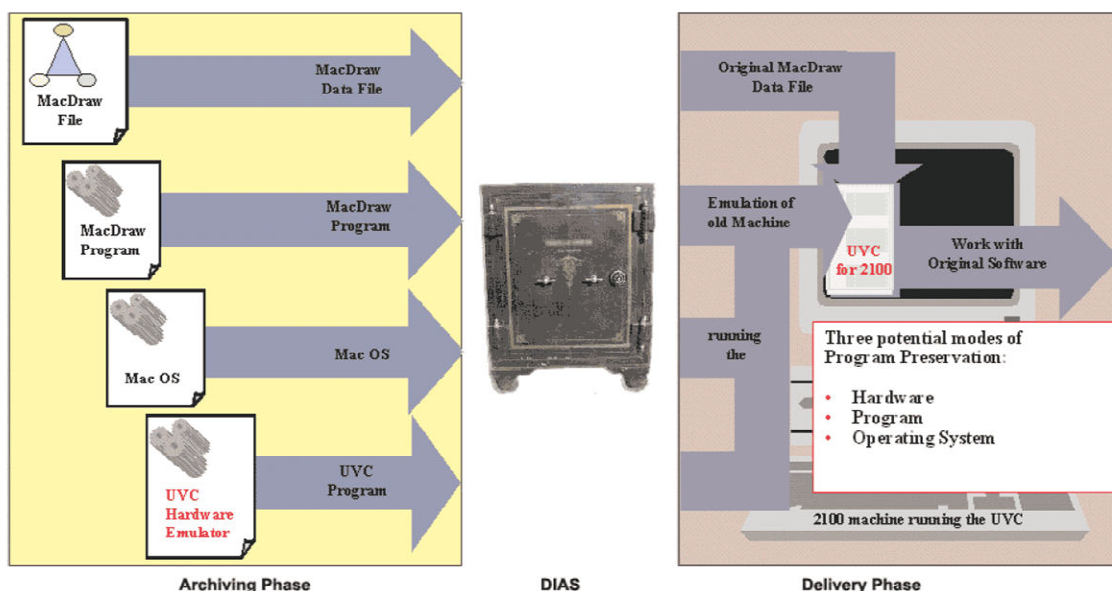


Figure 8 – IBM UVC DIAS preservation system.¹⁰¹

The IBM DIAS does not currently implement the UVC concept, but it has been ‘proofed’ using

¹⁰⁰ Dr Raymond J van Diessen and Dr. Johan F Steenbakkers, “The Long-term Preservation Study of the DNEP project” Page 28.

¹⁰¹ Dr Raymond J van Diessen and Dr. Johan F Steenbakkers, “The Long-term Preservation Study of the DNEP project” Page 31.

PDF as a test subject to IBM's own satisfaction. The dependence on writing the hardware emulators prior to hardware obsolescence introduces possible areas for failure, but overall it is a system worth considering for the hybrid solution, as it addresses both the issues of authenticity and of future access.

Instead of implementing the UVC system, the National Library of the Netherlands (KB) has implemented a 'Reference Platform' set of computers. These computers, set-up with a specific hardware specification, operating system and collection of applications, are able to access about 30 different file types which have been deemed suitable for the needs of the KB archive.

For now the IBM DIAS relies on current data backup technology, mainly magnetic tape and optical discs. In the sub-document to the IBM & KB DIAS study, "Managing media migration in a deposit system" they acknowledge that:

Most electronic deposit systems define their storage capacity needs in several TeraBytes (10^{12} Bytes). This requires media migration / refreshment to be consciously managed within the electronic deposit system to prevent a situation in which the time available for migration is insufficient for completing the process in time.¹⁰²

The situation described could easily arise given the volumes of raw data acquired daily by research institutions such as NASA. For them, the IBM DIAS solution will need to be combined with storage media, which will not degrade faster than the capacity of the archiving system to migrate or refresh the media. Any successful hybrid solution will have to take this factor into account.

¹⁰² Dr Raymond J van Diessen and Dr. Johan F Steenbakkens, "Managing media migration in a deposit system", IBM Netherlands, Amsterdam, December 2002

8. Conclusion

... the decree should be written on a stela of hard stone, in sacred writing, document writing, and Greek writing, and it should be set up in the first-class temples, the second-class temples and the third-class temples, next to the statue of the King, living forever.¹⁰³

The text above is taken from the final line of the Memphite Decree on the Rosetta Stone. It can be loosely interpreted as an instruction to preserve the contents of this document by translating it into three languages, inscribing it onto a very durable surface, and placing copies in at least three separate geographic locations.¹⁰⁴ This could be viewed as the first example of data preservation through an explicit archival strategy of redundancy!

The Rosetta Stone demonstrates fundamental principles in line with the archival issues outlined in Chapter 6 - Digital Archiving Issues. Longevity is ensured by choosing a durable storage medium. Access is ensured by providing three alternate translations of the text, thus raising the odds that at least one of the languages will still be spoken in the far future. The three languages also provide documentation about each other, in that they provide translations for each other. It could be argued that they were also used as backup and recovery, in that through the Rosetta Stone archaeologist were finally able to translate (restore knowledge of) Egyptian hieroglyphs. Data redundancy and reliability are supported by storing the physical objects at separate geographic locations. Authenticity is preserved as the display medium is a component of the preserved object. The total storage costs were met in the initial creation of the object.

The digital age requires a similar robust long-term solution. Documents of possible historic value are being 'born digital' and run the very real risk of disappearing to history simply through lack of an effective long-term preservation strategy. (Appendix # 8 - Other Historic Examples outlines two examples of important historic documents which would not have survived a digital era.)

There are several theoretical solutions which went beyond the scope of this thesis but are worth mentioning as they influenced the rationale implemented in outlining the 'digital problem'. In their paper "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital

¹⁰³ "The Rosetta Stone: translation of the demotic text", from the British Museum Website <http://www.thebritishmuseum.ac.uk/compass/ixbin/print?ENC917>, accessed 20/02/2004

¹⁰⁴ Steve Gilheany, "Permanent Digital Records and the PDF Format", ArchiveBuilders.com, 2000, available at "White Papers at ZDNet UK", <http://whitepapers.zdnet.co.uk/> accessed 12/08/2003

Documents”¹⁰⁵ Alan R. Heminger and Steven B. Robertson proposed a method of preserving access to digital documents. They suggest that:

... knowledge preserved about different storage devices and file formats can be used to recover data from obsolete media and to reconstruct the digital documents.¹⁰⁶

This supports the notion postulated by IBM that a UVC could be used to emulate not only software, but also complete OS and hardware systems. Steve Gilheany in his paper “Permanent Digital Records and the PDF Format”¹⁰⁷ also suggests a type of information wrapping, whereby a digital document gains layers of information as it progresses into the future, but the core of the document never changes. Users in the future use the layers of information to reverse engineer the interface necessary to view the document.

These proposed solutions are implementing the notion of documentation, as outlined in Chapter 6 - Digital Archiving Issues, to support access to digital documents in the long-term future. However they remain in the arena of the theoretical and do not yet provide a concrete, contemporary solution, which is a pressing concern. The pressing need for a current solution has been thoroughly demonstrated through the many examples of the ‘digital problem’ cited throughout this thesis. The solutions to be considered must be broken down into two major headings:

- A plan for the long-term.
- What can we do right now?

One more study worthy of mention is “Peer to peer data trading to preserve information”, by Brian Cooper and Hector Garcia-Molina¹⁰⁸. Cooper and Garcia-Molina propose that a network of archiving institutions could embark on a data trading system to ensure long-term preservation of information by maximising data redundancy. This highly technical document introduces the necessary trade deed concepts, trading algorithms, authenticity verifications, and simulations to demonstrate the most reliable policies. Their work influenced the concept in this thesis of grid computing as a viable hybrid solution component.

¹⁰⁵ Alan R. Heminger and Steven B. Robertson, “Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents”, 1998, European Research Consortium for Informatics and Mathematics (ERCIM) website, <http://www.ercim.org/publication/ws-proceedings/DELOS6/> accessed 16/03/2004

¹⁰⁶ Alan R. Heminger and Steven B. Robertson, “Digital Rosetta Stone” Abstract page 1

¹⁰⁷ Steve Gilheany, “Permanent Digital Records and the PDF Format”

¹⁰⁸ Brian Cooper and Hector Garcia-Molina, “Peer to peer data trading to preserve information”, Extended version, Department of Computer Science, Stanford University, February 22, 2001, <http://www-db.stanford.edu/~cooperb/pubs/trading.pdf> accessed 19/11/2003

8.1. A Plan for the Long-term

In the introduction to this thesis the concept of the 'chain of digital preservation' was outlined and then investigated in detail in chapters 3 through 5. It was demonstrated that a weak link in any area of the chain would lead to loss of access to the stored digital information, and many examples were cited to show that this is a serious and pressing issue.

The requirements for long-term digital preservation were investigated in detail through Chapter 6 - Digital Archiving Issues so that the range of needs would be fully grasped. Any proposed long-term solution must address a majority if not all of these requirements.

In Chapter 7 - Digital Archiving Solutions, many possibilities were explored and the strengths and / or weaknesses highlighted. It was mooted that a hybrid solution would successfully combine the best aspects of each to resolve a truly long-term solution to the 'digital problem'. A hybrid solution would implement differing measures for differing needs. Documents and data currently judged as having historic value can be moved to a static and robust storage medium that fulfils most of the archival requirements. 'Live' documents which may be required for ongoing access or have not been yet judged as to their future value can be maintained through a more fluid and consequently less guaranteed process which will develop and change over time.

Solutions such as those posited by the Eastman Kodak Company and by Norsam Technologies involve a static, time capsule approach to long-term preservation, which ignore such aspects as authenticity and, for the most part, ongoing access. However, Norsam Technologies in particular presents a very robust storage media. What is surprising is that Norsam has not developed an equivalent to optical disc using their technology.

Imagine using the micro-tooling technology to create an extremely durable albeit highly expensive DVD. This would produce a long-term digital storage medium, which would meet several more of the criteria wished for, such as immediate access (digital, as opposed to the analog use proposed by Norsam), authenticity, and the ability to withstand long periods of dormancy. It would not inherently solve the many problems associated with obsolescence in its many forms, mostly software and hardware, but it would be a step in the right direction.

Micro-tooling would be a write-once storage solution, as opposed to re-writable storage media, such as magnetic disks or optical RW discs. Therefore this type of technology may not be appropriate for ongoing storage requirements, such as backing up networked data. Rather micro-tooling technology would continue in its present form as a static, time-capsule approach to long-term storage, most appropriate perhaps to large institutional collections.

Micro-tooling technology could be a long-term and sufficiently storage intense solution compared to holographic memory, which may suffer the same unpredictable life expectancies as other light-sensitive media. Whether or not a home PC micro-tooling solution is realistic is another issue, as requirements include a vacuum environment, use of liquid metals and gases to produce the ion beam and achieve the milling procedure, etc.¹⁰⁹

Another ingredient for the hybrid solution was introduced in Chapter 7 - Digital Archiving Solutions via the existing technologies of SAN, NAS, and grid computing. This proposes that an ongoing and transparent (hiding complexity) process of refreshing digital data be managed via a network of storage systems, based on the RAID concept of redundancy of data and hardware. This system would address the issues of daily preservation which are encountered because of the unreliable nature of current storage media, as was shown in chapters 3 - Storage Media Degradation and 4 - Hardware Obsolescence. By shifting storage from unreliable removable storage to a networked system of storage resources, a much more robust and dependable environment can be accessed by individual users.

This grid computing system of storage is not currently available to the average computer user, but it is a serious consideration for development in the near future. It was shown in Chapter 7 - Digital Archiving Solutions that a large number of research organisations and businesses are already investing significant time and resources in this branch of technological development. The main missing ingredient for users is the lack of broadband connections, and any grid computing solution will remain stifled until broadband access becomes commonplace.

The third ingredient for the hybrid solution introduced was the IBM Digital Information Archiving System (DIAS) which implemented an emulation concept called the Universal Virtual Computer (UVC). This radical conceptual model allows documents stored for the long-term to overcome the problems of authenticity, software obsolescence and hardware obsolescence. This could be utilised in conjunction with either a static or a live digital storage solution.

The only drawback apparent from these systems is that they will become difficult to access after long periods of dormancy, due to possible losses of documentation in the interim period. The IBM DIAS UVC system is dependent upon an understanding of the UVC concept, after which emulating any historic system archived should be fairly straight forward. However this could be said of traditional archaeological finds as well, that the intervening period of time measured in centuries or millennia make it difficult to interpret information stored in found objects.

¹⁰⁹ Delft Institute of Microelectronics and Submicron Technology website, <http://dsa.dimes.tudelft.nl/usage/technology/FIB/> accessed 25/02/2004

The micro-tooling concept runs the risk of hardware obsolescence, as does any new technology in this 'transition period' of developing information technology.

8.2. What can we do right now?

The hybrid solutions proposed are some years if not decades away from any type of widespread availability. In the meantime computer users urgently require a real world solution to the 'digital problem'. The preservation of digital cultural heritage could turn out to be the current generation's greatest challenge.

The individuals and small businesses which produce between 10MB and 10GB of data yearly must rely on their own vigilance to ensure that stored information does not fall prey to any of the weak links in the chain of digital preservation.

Institutions and archives face an ever increasing storage requirement as their collections of information grow on a daily basis. Automated systems employing SAN technology and backing up to tape (still the lowest cost by volume) will have to be set up in accordance with the considerations for long-term digital preservation introduced in Chapter 6 - Digital Archiving Issues. Inevitably at some point in the future the amount of information stored will exceed the institutions' ability to refresh to newer media within the given life span of the media.¹¹⁰ This will be overcome by adding more refreshing hardware, doubling up to spread the processing load, but this will not be an optimal storage solution as it will only increase expenses and hardware maintenance considerations.

Other institutions may opt for hybrid solutions employing microfilm as the long-term storage media for 'relatively static' digital documents. The benefits of this tested archiving technology may in the short term outweigh the losses of authenticity and digital access, however it should not be construed as a digital solution, as it ignores the many digital media formats that go beyond static text. This is also an expensive option compared to digital storage, as digital storage media currently costs a fraction of microfilm and paper per megabyte of information.¹¹¹

Perhaps the next storage technology on the horizon, holographic memory for example, will prove to be robust, dependable, and become a widely accepted de-facto standard. It would appear that up to this point that is what many investors in digital technology (practically every modern business and institution on the planet) have been pinning their hopes on, a science fiction type answer to all their storage needs.

¹¹⁰ Dr Raymond J van Diessen and Dr. Johan F Steenbakkens, "Managing media migration in a deposit system"

¹¹¹ See Appendix # 7 - Document Storage Costs

The current 'hybrid solution' will need to be composed of a mix of equal parts awareness and vigilance. The immediate solution must entail educating the general computer user as to the true situation regarding storage of any form of digital information. As outlined in this thesis the situation is tenuous at best.

The lifetime of digital information is dictated by many factors, as has been clearly demonstrated in this thesis. These factors include 'storage media degradation', 'hardware obsolescence', and 'software or file format obsolescence'. Each of these factors must be anticipated and acted upon in a timely fashion. Vigilance entails that the archive always be aware of the longevity status of any given digital document in the collection. An understanding of the limitations of storage media will go a long way to encouraging a stricter data management routine.

Appendix # 1 - Definitions

Definitions supplied by the author, unless stated otherwise.

Archiving refers to the entire range of considerations necessary to successfully preserve digital information in the long-term.

ASCII - American Standard Code for Information Interchange: This is the de facto world-wide standard for the code numbers used by computers to represent all the upper and lower-case Latin letters, numbers, punctuation, etc. There are 128 standard ASCII codes, each of which can be represented by a 7 digit binary number. The inability of US-ASCII to represent nearly any language other than English lead to the development of international extensions to US-ASCII, such as Latin-1. *Definition by Hyperdictionary.com, <http://www.hyperdictionary.com> accessed 08/03/2004*

'Born Digital' refers to documents that are created on the computer, as opposed to ones that have been digitised by being scanned.

'Browser wars' is a term commonly used to refer to the ongoing competition between Netscape Navigator and Microsoft's Internet Explorer and several other internet browser software packages.

Checksum: a digit representing the sum of the digits in an instance of digital data, and is used to check whether errors have occurred in transmission or storage. A computed value which depends on the contents of a block of data and which is transmitted or stored along with the data in order to detect corruption of the data. The receiving system recomputes the checksum based upon the received data and compares this value with the one sent with the data. If the two values are the same, the receiver has some confidence that the data was received correctly. *Definition by Hyperdictionary.com, <http://www.hyperdictionary.com> accessed 08/03/2004*

CODEC - Acronym for coder-decoder. An electronic device, circuit, or software that converts digital signals to and from analog. Usually also includes digital compression technology for added efficiency. Different codecs may provide different efficiency, quality, and features. *Definition by Hyperdictionary.com, <http://www.hyperdictionary.com> accessed 08/03/2004*

The term **Component** in relation to digital technology can refer to hardware, software or data. A component is a fully developed unit which can be re-used in varying situations. A software component is a unit of code that completes one task, and can be slotted into larger programmes. A hardware component is a replaceable part which can be integrated into any system designed to receive it. Data components contain content that can be accessed and manipulated by applications. *Definition adapted from the Open Process Framework website, <http://www.donald-firesmith.com/>, and the WikiWikiWeb discussion on Component Definitions, <http://c2.com/cgi/wiki?ComponentDefinition>. Accessed 26/04/2004*

Data Conversion involves changing the digital file format to another format, such as converting from a Word file to a PDF document. Conversion almost always involves the loss or corruption of information.

Data Migration refers to the transfer of digital signals from one storage media type to another, for example from magnetic tape to CD.

Data Refreshing involves transferring digital signals to an identical but newer storage media, i.e. tape to tape, CD to CD, etc.

International Standards Organisation (ISO) – a world-wide federation of national standards bodies from some 100 countries, one from each country. ISO is a non-governmental organisation established in 1947. The mission of ISO is to promote the development of standardisation and related activities in the world with a view to facilitating the international exchange of goods and services, and to developing co-operation in the spheres of intellectual, scientific, technological and economic activity. ISO's work results in international agreements which are published as International Standards. ISO website <http://www.iso.ch/> accessed 11/03/2004

Long-term Digital Preservation: Continued access to digital materials, or at least to the information contained in them, indefinitely, i.e. beyond the limits of media failure or technological change. *Definition adapted from The Digital Preservation Coalition, <http://www.dpconline.org/>, accessed 27/02/04*

Media Degradation refers to the actual object that the digital signal is stored on, such as the floppy disk or magnetic tape.

Mirroring - Writing duplicate data to more than one device, in order to protect against loss of data in the event of device failure. *Definition adapted from Hyperdictionary.com, <http://www.hyperdictionary.com> 09/03/2004*

Obsolescence - the process of becoming obsolete; falling into disuse or becoming out of date; "a policy of planned obsolescence" *Definition by Hyperdictionary.com, <http://www.hyperdictionary.com> accessed 08/03/2004*

Obsolete - no longer in use; "obsolete words"; old; no longer in use or valid or fashionable; "obsolete words"; "an obsolete locomotive"; "outdated equipment"; "superannuated laws"; "out-of-date ideas" *Definition by Hyperdictionary.com, <http://www.hyperdictionary.com> accessed 08/03/2004*

Parity - an error detection procedure in which a bit (0 or 1) is added to each group of bits so that it will have either an odd number of 1's or an even number of 1's. If the parity is odd then any group of bits that arrives with an even number of 1's must contain an error. *Definition adapted from Hyperdictionary.com, <http://www.hyperdictionary.com> 09/03/2004*

Striping – A process whereby segments of data can be written to multiple physical devices in a round-robin fashion. This is useful if the processor is capable of reading or writing data faster than a single disk can supply or accept it. *Definition adapted from Hyperdictionary.com, <http://www.hyperdictionary.com> 09/03/2004*

World Wide Web Consortium (W3C) – develops interoperable technologies (specifications, guidelines, software, and tools) to lead the Web to its full potential. W3C is a forum for information, commerce, communication, and collective understanding. As of 11/03/2004 the W3C has 371 members (companies). W3C website <http://www.w3.org/> accessed 11/03/2004

Appendix # 2 - Abbreviations

AIP	Archival Information Package
ASCII	American Standard Code for Information Interchange
BIOS	Basic Input Output System
CLIR	Council on Library and Information Resources
CSS	Cascading Style Sheets
DIAS	Digital Information Archiving System
DOS	Disk Operating System
EDAP	Extended Data Availability & Protection
HTML	Hyper Text Markup Language
ICT	Information and Communication Technology
IE	(Microsoft) Internet Explorer
ISO	The International Standards Organisation
JPEG	Joint Photographer's Expert Group
KB	Koninklijke Bibliotheek – The National Library of the Netherlands
LAN	Local Area Network
MTTF	Mean Time To Failure
NAS	Network Attached Storage
NIST	National Institute of Standards and Technology
OS	Operating System
PDL	Page Description Language (PostScript)
PMCIA	Personal Computer Memory Card International Association. (or People Can't Memorise Computer Industry Acronyms)
RAB	RAID Advisory Board
RAID	Redundant Array of Independent (or Inexpensive) Drives
SAN	Storage Area Network
UVC	Universal Virtual Computer
W3C	The World Wide Web Consortium
W3C	World Wide Web Consortium
WAN	Wide Area Network
XHTML	eXtensible Hyper Text Markup Language
XML	eXtensible Markup Language

Appendix # 3 - Farewell My Floppy

'Farewell my floppy: A strategy for migration of digital information.'

Deborah Woodyard, Electronic Preservation, National Library of Australia

Summary table of data transfer test conducted by Deborah Woodyard, 1996

Initial Selection - 64 Items (Floppy disks)			
	Action taken	Results / Problems	# Items in process
1.	Hardware, OS and disc drive	23 items could not be used due to lack of proper hardware or software. 1 item was damaged	42 items remain
2.	Virus check	No viruses detected	42 items remain
3.	Scan for media deterioration	2 discs were faulty 1 duplicate was located.	41 items remain
4.	Run from floppy	1 disc discovered as blank 13 publications could not be run due to lack of hardware or software	27 items remain
5.	Copy contents to network drive	Successful	40 items copied
6.	Compare original & copy	Successful	40 items confirmed
7.	Run from network drive	4 items required to run from A: drive only	23 items run
8.	Apply ISO 9660 naming conventions	Many items would not function properly on renaming (exact amount not listed)	?
9.	Create ASCII text documents to accompany each item with spec notes (metadata). Text document includes following information: hardware required, software required, notes to assist installation, number of files, storage size (Mb), original format and number of disks, existence of accompanying materials, content description (e.g. text, database, software), any copyright statement from the publisher, date of transfer to new format and statement of format type (for future transfer reference)		40 items
10.	Write to CD-R	4 file names rejected for containing illegal characters	40 items + ASCII text (~1,900 files)
11.	Compare original & copy	Successful	
12.	Run from CD-R	5 items would not function 12 items did not have required software 1 item could not be tested	22 items run

Appendix # 4 - Basic Layers of CDs and DVDs

Byers, Fred R. "Care and Handling of CDs and DVDs - A Guide for Librarians and Archivists", NIST Special Publication 500-252, NIST & CLIR, USA, 2003

Basic Layers of CDs and DVDs				
Basic layers of CD-ROM and DVD-ROM (Replicated discs for audio, video, computer use, or interactive games)				
CD-ROM (Single-sided)	DVD-ROM (Single-sided)	DVD-ROM (Single-sided)	DVD-ROM (Double-sided)	DVD-ROM (Double-sided)
(All CD-ROMs are one-sided) One recorded layer	(One side) One recorded layer	(One side) Two recorded layers	(Both sides) One recorded layer per side	(Both sides) Two recorded layers per side
Label, optional	Label, optional	Label, optional	Label, optional (hub area only)	Label, optional (hub area only)
Lacquer	Polycarbonate	Polycarbonate	Polycarbonate	Polycarbonate
Metal	Center adhesive	Metal (fully-reflective)	Metal	Metal (semi-reflective)
Polycarbonate	Metal	Center adhesive	Center adhesive	Adhesive
	Polycarbonate	Metal (semi-reflective)	Metal	Metal (fully-reflective)
		Polycarbonate	Polycarbonate	Center adhesive
			Label, optional (hub area only)	Metal (fully-reflective)
				Adhesive
				Metal (semi-reflective)
				Polycarbonate
				Label, optional (hub area only)
Basic layers of CD -R/-RW and DVD -R/-RW/+R/+RW/RAM (Blank recordable discs for all applications listed for ROM discs)				
CD-R, CD-RW (Single sided)	DVD-R, DVD-RW, DVD+R, DVD+RW, DVD-RAM (Single sided)	DVD-R, DVD-RW, DVD+R, DVD+RW, DVD-RAM (Double sided)		
CD-R/RW are one-sided, One recordable layer only	(One side) One recordable layer only	(Both sides) One recordable layer per side only		
Label, optional	Label, optional	Label, optional (hub area only)		
Lacquer	Polycarbonate	Polycarbonate		
Metal	Center adhesive	Recording/writing layer		
Recording/writing layer	Metal	Metal		
Polycarbonate	Recording/writing layer	Center adhesive		
	Polycarbonate	Metal		
		Recording/writing layer		
		Polycarbonate		
		Label, optional (hub area only)		

Appendix # 5 - Definitions of RAID levels

Adapted from the RaidWeb website
<http://www.raidweb.com/whatis.html> accessed 09/03/2004

RAID Level	Common Name	Description	Array's Capacity	Data Reliability	Data Transfer Capacity	Minimum Drive Required
0	Disk striping	Data distributed across the disks in the array. No redundant information provided.	(N) disks	Low	Very High	2
1	Disk mirroring	All data duplicated	1* disks	Very High	High	2
3	Parallel transfer disks with parity	Data sector is subdivided and distributed across all data disk. Redundant information stored on a dedicated parity disk.	(N-1) disks	Very High	Highest of all listed alter-natives	3
5	Independent access Array without rotating parity	Data sectors are distributed as with disk striping, redundant information is interspersed with user data.	(N-1) disks	Very High	Very High	3
0+1	Disk-Striping + Disk-Mirroring	Combined striping and mirroring function without parity. Fast data access and single drive fault tolerance.	(N/2) disks	Very High	High	4

Appendix # 6 - Computer storage timeline

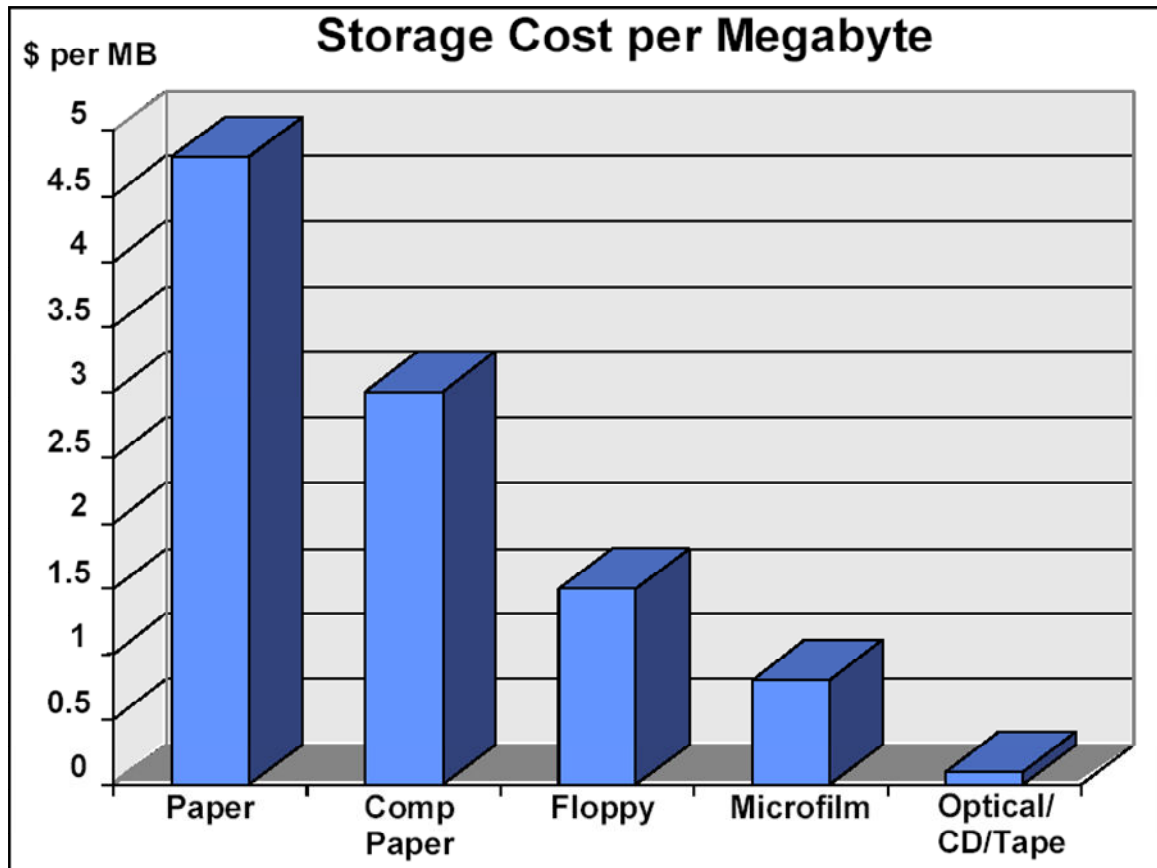
Adapted from PCmuseum website's "Hard Disk Drive History"

<http://www.fortunecity.com/marina/reach/435/storage.html> accessed 25/03/2004

Removable storage	Non-removable storage
1940	Vacuum tubes and punch cards.
1950	Introduction of tape drives
1950s	mercury delay lines, magnetic drums, electrostatic storage tubes and magnetic core storage
1956	The IBM® 305 RAMAC (Random Access Method of Accounting and Control) is the first magnetic hard disk for data storage. It required 50 24-inch disks to store 5MB of data.
1967	IBM® builds the first floppy disk .
1971	IBM® releases the 8-inch floppy plastic disk coated with iron oxide.
1973	IBM® introduces the IBM® 3340 Winchester hard disk unit . The recording head rides on a layer of air 18 millionths of an inch thick.
1976 - AUG.	Shugart announces its 5.25 inch "minifloppy" disk drive for US\$390.
1980	Sony Electronics introduces the 3.5 inch floppy disk drive , double-sided, double-density, holding up to 875KB unformatted.
1980 - JUN.	Seagate Technologies announces the first 5.25-inch hard disk drive .
1982 - SEP.	Iomega begins production of a 10MB 8-inch floppy-disk drive.
1982 - NOV.	Drivetec announces the Superminifloppy, 3.33MB unformatted on a 5.25-inch drive.
1982 - DEC.	Amdek releases the Amdisk-3 Micro-Floppy-disk Cartridge system - two 3-inch floppy drives
1982	Davong Systems introduces its 5MB Winchester Hard Disk Drive for the IBM® PC.
1983 - MAY.	Sony Electronics increases the storage capacity of the 3.5 inch floppy disk to 1MB.
1983	With the introduction of the IBM® PC/XT , hard disk drives became a standard component of most personal computers.
1983 -	Philips and Sony develop the CD-ROM , as an extension of audio CD technology.
By 1987	3.5-inch hard drives began to appear. This format quickly became the standard for desktop and portable systems requiring less than 500 MB capacity.
1987 - SEP.	Microsoft ships Microsoft Bookshelf, its first CD-ROM application.
1990 - JAN.	Commodore proposes an "interactive graphics player", based on a variant of the Amiga 500, with 1MB of RAM. The machine includes an integrated CD-ROM drive , but no keyboard.
1991 - OCT.	Insite Technology begins shipping its 21 MB 3.5-inch floppy disk drive to system vendors. The drive uses " floptical " disks, using optical technology to store data.
By 1992	A number of 1.8-inch hard drives appeared delivering capacities up to 40 MB. Even a 1.3-inch hard drive , about the size of a matchbox, was introduced.
1993 - OCT.	NEC Technologies unveils the first triple-speed (450KBps) CD-ROM drive .
1994 - JAN.	NEC Technologies ships its quad-speed CD-ROM , priced at US\$1000.
1994 - DEC.	Iomega Corp. introduces its Zip drive and disks , removable storage in sizes of 25MB or 100MB.
1997 - JUNE	Imation release the SuperDisk 100MB diskettes and drives.
1997 - NOV.	IBM® releases the Deskstar 16GP, a 16.8-gigabyte hard drive . This brings down the cost of storage to .25 cents per megabyte.
1998 - NOV.	IBM® announced a 25GB hard drive. That first hard disk drive in 1956 had a capacity of 5 megabytes. IBM's Deskstar 25GP 25GB drive has 5,000 times the capacity of that first drive.
1999 - OCT.	IBM® releases the 10,000 RPM Ultrastar 72ZX -- the world's highest capacity drive at 73GB.
2000 - JUNE	IBM® announced the availability of the 1Gb Microdrive, the world's smallest, lightest and largest capacity mobile hard disk increasing storage by a factor of three.
2003 - MAY	Imation have discontinued the manufacture of Imation SuperDisk™ 120MB diskettes and drives.

Appendix # 7 - Document Storage Costs

From "Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative"
Intergraph Solutions Group, Madison AL, 2003
Page 5



Appendix # 8 - Other Historic Examples

Vincent Van Gogh

Over the course of 18 years Vincent Van Gogh wrote over nine hundred letters to his brother Theo.

“He wrote the way other people keep a diary, with news about everyday events, comments on books, art and artists, revelations about his expectations of life, and his fears about illness and death. ... An important part of the letters consists of passages in which Van Gogh describes the drawings and paintings he was working on at the time, and what it was that made him choose a specific subject.”¹¹²

After his death in July 1890, and his brother Theo's death six months later, Theo's wife Johanna returned to Holland where she actively promoted his artwork. By 1914 his work had gained international recognition and so Johanna Van Gogh-Bonger published the correspondence between the two brothers in a three-volume set. Further editions have regularly been published since, expanding on the original collection.

Consider the fate of this correspondence had it been carried out via email.

Sumerian Clay Cuneiform Tablets

Sumerian clay cuneiform tablets from 5,200 years ago (3,200 BC) contain ancient texts that vary from an original “Garden of Eden” type story (1,800 BC), to school lessons, king lists, astronomical records, hymns, epic tales, and humdrum sets of business accounts and contracts. An ever expanding Sumerian Dictionary is being published online at <http://ccat.sas.upenn.edu/psd> in an attempt to combine various researchers' knowledge into one document. Up until this project, students have had to compile their own specialised dictionaries before they could begin any translation work, due to the lack of a centralised resource of Sumerian translations.

“Sumerian is a very difficult and obtuse language” says Tinney*. “It has no relatives, living or dead. The script is not syllabic, let alone alphabetic - it is logographic, like Chinese.”¹¹³

This most ancient of written languages predates Egyptian hieroglyphics by several hundred years. Thousands of these tablets have been recovered (60,000 in 1990 alone), but only about 1 percent of them have ever been read. The possible importance of their translations cannot be underestimated, as there are more cuneiform tablets in existence today than there are medieval manuscripts, even though the tablets are nearly 10 times older.

None of today's digital storage media are designed to last for 5,200 years.

¹¹² **Van Gogh Museum Guide**, Van Gogh Museum, Amsterdam, 1996

* Dr. Steve Tinney, curator of the tablet collection in the Museum of Archeology and Anthropology, University of Pennsylvania in Philadelphia. He is co-editor of the Sumerian Dictionary Project with Sumerologist Åke Sjöberg.

¹¹³ Clark, Arthur. **Sumerians on the Information Superhighway**, Aramco World, Vol 51, No.2, March/April 2000, Aramco Services Company, USA.

Bibliography

[This is a list of the sources that I found useful in composing this thesis. Not all of these are directly cited in the text.]

Primary Sources

Brian Cooper and Hector Garcia-Molina, "Peer to peer data trading to preserve information", Extended version, Department of Computer Science, Stanford University, February 22, 2001, <http://www-db.stanford.edu/~cooperb/pubs/trading.pdf> accessed 19/11/2003

Byers, Fred R. "Care and Handling of CDs and DVDs - A Guide for Librarians and Archivists", CLIR & NIST Special Publication 500-252, NIST & CLIR, USA, 2003
<http://www.itl.nist.gov/div895/carefordisc/CDandDVDCareandHandlingGuide.pdf> accessed 19/02/2004

David A. Patterson, Garth A. Gibson and Randy H. Katz. "A Case for Redundant Arrays of Inexpensive Disks (RAID)" SIGMOD Conference 1988

Dr Raymond J van Diessen and Dr. Johan F Steenbakkens, "Managing media migration in a deposit system", IBM Netherlands, Amsterdam, December 2002

Dr Raymond J van Diessen and Dr. Johan F Steenbakkens, "The Long-term Preservation Study of the DNEP project - an overview of the results", IBM Netherlands, Amsterdam, December 2002

Gilheany, Steve. "Permanent Digital Records and the PDF Format", ArchiveBuilders.com, 2000, White Papers at ZDNet UK, <http://whitepapers.zdnet.co.uk/> accessed 12/08/2003

Harvey, Ross, 1995 NPO Conference Paper, National Library of Australia website, <http://www.nla.gov.au/niac/meetings/np095rh.html#roth> accessed 27/01/2004

Hedstrom, Margaret, "It's About Time: Research Challenges In Digital Archiving And Long-Term Preservation", Final Report: Workshop On Research Challenges In Digital Archiving And Long-Term Preservation, The National Science Foundation & The Library of Congress, 2002, <http://www.digitalpreservation.gov/index.php?nav=3&subnav=11> accessed 10/08/2003

Lawrence, H. Andrew. "Digital Insurance For Information At Risk. A strategic overview of digital preservation", Eastman Kodak Company, 2000

Samuel D. Gasster, Craig D. Lee, Brooks Davis, Matt Clark, Mike AnYeung, John R. Wilson, Debra M. Ladwig. "The NASA/GSFC Advanced Data Grid: A Prototype for Future Earth Sciences Ground System Architectures", NASA / GSFC, March 2003, <http://sunset.usc.edu/gsaw/gsaw2003/s7/gasster.pdf> accessed 16/03/2004

Seamus Ross and Ann Gow. "Digital Archaeology: Rescuing Neglected and Damaged Data Resources", A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials, Humanities Advanced Technology and Information Institute (HATII), University of Glasgow, February 1999, <http://www.hatii.arts.gla.ac.uk/>, accessed 23/02/04

Van Bogart, John. "Magnetic Tape and Handling: A Guide for Libraries and Archives", Washington DC: The Commission on Preservation and Access and National Media Laboratory, 1995; http://www.imation.com/en_US/main.jhtml?Id=64_04_02 accessed 23/02/2004

Woodyard, Deborah. "Data Recovery and Providing Access to Digital Manuscripts", National Library of Australia, 2001 <http://www.nla.gov.au/nla/staffpaper/woodyard3.html> accessed 5/03/2004

Woodyard, Deborah. "Farewell My Floppy: A strategy for migration of digital information", National Library of Australia, 1997 <http://www.nla.gov.au/nla/staffpaper/valadw.html> accessed 23/02/2004

Other Important Sources

"Data Integration, Interoperability, and Conversion Services for US Army Corps of Engineers Automated Document Conversion Strategy Initiative", Intergraph Solutions Group, Madison AL, 2003 http://tsc.wes.army.mil/downloads/ADCS_Final_Report_Main.pdf accessed 23/02/2004

"Digital Preservation Testbed: From digital volatility to digital permanence - Preserving email", The Hague, April 2003 <http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-email-en.pdf> accessed 03/03/2004

"Fend off data degradation", Quality Online, May 1999 <http://www.jdhunt.com/QualityMag1.pdf> accessed 28/11/2003

"Protecting Knowledge Assets", IT Storage Online, published 1/31/2003, <http://www.itstorageonline.com/contents/news/> accessed 12/08/2003.

"The Internet Under Crisis Conditions: Learning from September 11", National Academy Press, <http://www.nap.edu/>, March 2003, accessed 04/12/2003

"White House E-Mail: The top-secret computer messages the Reagan / Bush White House tried to destroy", edited by Tom Blanton, National Security Archives, Washington, 1995.

Accurite Technologies, "Floppy Disk Drive Primer", <http://www accurite.com/FloppyPrimer.html> accessed 23/02/2004

Alan R. Heminger and Steven B. Robertson, "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-term Access to Digital Documents", 1998, The European Research Consortium for Informatics and Mathematics (ERCIM) website <http://www.ercim.org/publication/ws-proceedings/DELOS6/> accessed 16/03/2004

Booyens, Vic. "Information and the Archiving thereof", Enterprise Storage Comparex Africa (Pty) Ltd., 2003, The Saice IT website http://www.saice-it.co.za/pdf/24th/Vic_Booyens_Paper_Information_And_The_Archiving.pdf accessed 09/03/2004

Bradbury, Danny. "Virtually the next big thing", Computer Weekly, October 10 2002, http://www.findarticles.com/cf_0/m0COW/2002/_Oct_10/92671444/print.jhtml accessed 10/03/2004

Brian Cooper and Hector Garcia-Molina, "InfoMonitor: Unobtrusively archiving a World Wide Web server", Department of Computer Science, Stanford University, 2000, <http://citeseer.ist.psu.edu/cooper01infomonitor.html>, 19/11/2003

Brian Cooper, Arturo Crespo and Hector Garcia-Molina, "Implementing a Reliable Digital Object Archive", Department of Computer Science, Stanford University, 1999, <http://citeseer.ist.psu.edu/264311.html> accessed 19/11/2003

Brian F. Cooper, Arturo Crespo and Hector Garcia-Molina, "The Stanford Archival Repository Project: Preserving our digital past", Department of Computer Science, Stanford University, 2003, <http://www-db.stanford.edu/~crespo/publications/lirn.pdf> accessed 19/11/2003

Bricklin, Dan website <http://www.visicalc.org/history/vcexecutable.htm> accessed 11/03/2004

Carr, Nicholas G. "IT Doesn't Matter", Harvard Business Review, Reprint R0305B, May 2003

Clark, Aurthur. "Sumerians on the Information Superhighway", Aramco World, Vol. 51, No.2, March/April 2000, Aramco Services Company, USA.

Delft Institute of Microelectronics and Submicron Technology website, <http://dsa.dimes.tudelft.nl/usage/technology/FIB/> accessed 25/02/2004

Faughan, Leslie. "Public sector is ready for disaster", Digital Ireland, The Irish Independent Newspaper, February 2004.

Finney, Andy. The Domesday Project website, <http://www.atsf.co.uk/dottext/domesday.html> accessed 02/03/2004

Fischer, Paul, "Considerations for Building Long-life OEM Systems", White Paper, Radisys Corporation, September 1999, <http://www.radisys.com/files/07-1068-00.pdf> accessed 11/12/2003

Flink, Chuck. "Falling on their Face: Six Incidents from Corporate History", May 27th, 2000 http://www.activewin.com/editorials/charles_flink/ink/23.shtml accessed 12/03/2004

Grid Café website, "Grid projects in the world", <http://gridcafe.web.cern.ch/gridcafe/gridprojects/athome.html> accessed 15/03/2004

Grid Café website, "What is the grid?", <http://gridcafe.web.cern.ch/gridcafe/whatisgrid/reality.html> accessed 15/03/2004

Howstuffworks.com website, "How Flash Memory Works", <http://computer.howstuffworks.com/flash-memory.htm> accessed 24/02/2004

Howstuffworks.com website, "How Holographic Memory Will Work", <http://computer.howstuffworks.com/holographic-memory.htm>, accessed 24/02/2004

Howstuffworks.com, "How 3-D Graphics Work", <http://computer.howstuffworks.com/3dgraphics8.htm> accessed 26/04/2004

IBM website, "What is grid computing", http://www-1.ibm.com/grid/about_grid/what_is.shtml accessed 03/03/2004

Imation website, "Imation SuperDisk Technology, Keeping You Informed" http://www.imation.com/en_US/product.jhtml?Id=IM_FAM122 accessed 10/03/2004

Iomega website, "Iomega® Legacy Products: Discontinued Data Storage Drives from Iomega", http://www.iomega.com/na/products/product_family.jsp accessed 12/03/2004

James H. Cowie, Andy T. Ogielski, BJ Premore, Eric A. Smith and Todd Underwood, "Impact of the 2003 Blackouts on Internet Communications", Preliminary Report, Renesys Corporation, November 21 2003

Jason G. Cummins & Giovanni L. Violante, "Investment-Specific Technical Change in the US (1947–2000): Measurement and Macroeconomic Consequences", Federal Reserve Board & University College London, CEPR, January 16, 2002

John Ward-Perkins and Amanda Claridge, "Pompeii AD 79", Westerham Press, England, 1976

Keegan, Victor, "Erasing the information age", The Guardian, January 10 2002, <http://www.guardian.co.uk/> accessed 27/02/2004

Kidman, Angus. "Storage: The Inside Story", Technology & Business Magazine, 11 September 2002, <http://www.zdnet.com/printfriendly?AT=2000023527-20268129> accessed 28/11/2003

Kridler, Chris. "Digital history is vanishing", Florida Today, February 25, 2001. <http://www.floridatoday.com/news/people/stories/2001/feb/peo022501a.html> accessed 20/02/2004

Levy, Steven, MSNBC 2004 Newsweek article "OK, Mac, Make a wish", <http://msnbc.msn.com/id/4052227> accessed 12/03/2004

Levy, Steven. "Hello Again", Newsweek 1998
<http://www.geocities.com/ResearchTriangle/2952/imacnews.html> accessed 12/03/2004

Microsoft website, "Microsoft Product Lifecycle Dates",
[http://support.microsoft.com/default.aspx?scid=fh;\[In\];LifeWin](http://support.microsoft.com/default.aspx?scid=fh;[In];LifeWin) accessed 11/03/2004

Norsam Technologies website, <http://www.norsam.com/rosetta.html>, accessed 20/02/2004

Ontrack Data Recovery Europe Ltd, "Understanding Data Loss",
<http://www.ontrack.com/understandingdataloss/> accessed 23/02/2004

Pockley, Simon. "Killing the Duck to Keep the Quack",
<http://www.acmi.net.au/FOD/FOD0055.html> accessed 20/02/2004

Ross J. Anderson, Vaclav Matyas Jr., Fabien A.P. Petitcolas, "The Eternal Resource Locator: An Alternative Means of Establishing Trust on the World Wide Web" University of Cambridge Computer Laboratory, 1998, <http://www.petitcolas.net/fabien/publications/ec98-erl.pdf> accessed 09/12/2003

Rothenberg, Jeff, "An Experiment in Using Emulation to Preserve Digital Publications", RAND-Europe, The Koninklijke Bibliotheek Den Haag, April 2000
<http://citeseer.ist.psu.edu/rothenberg00using.html> accessed 09/12/2003

Rothenberg, Jeff, "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation", Council on Library and Information Resources, Washington, DC, 1999,
<http://citeseer.ist.psu.edu/288848.html> accessed 09/12/2003

Rothenberg, Jeff, "Digital Information Lasts Forever – Or Five Years, Whichever Comes First", n.p., 2001, <http://www.amibusiness.com/dps/rothenberg-arma.pdf> accessed 09/12/2003

Schofield, Jack. "Digital dark age looms", The Guardian, January 9, 2003,
<http://www.guardian.co.uk/online/story/0,3605,871091,00.html> accessed 20/02/2004

Stepanek, Marcia, "Data Storage: From Digits to Dust", Business Week 20/04/1998,
<http://www.businessweek.com/archives/1998/b3574124.arc.htm> accessed 20/02/2004

The British Museum Website, "The Rosetta Stone: translation of the demotic text",
<http://www.thebritishmuseum.ac.uk/compass/ixbin/print?ENC917>, accessed 20/02/2004

UNESCO, "Charter on the Preservation of Digital Heritage", 15 October 2003,
http://portal.unesco.org/en/ev.php@URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html accessed 12/03/2004

USByte.com website, "RAID Systems" http://www.usbyte.com/common/raid_systems.htm
 accessed 09/03/2004

Van Gogh Museum Guide, Van Gogh Museum, Amsterdam, 1996

Vedantam, Shankar. "Space Shuttles Bound to Technologies of the Past", The Washington Post, February 25, 2003; <http://www.vedantam.com/obsolescence02-2003.html> accessed 19/11/2003

Waters, Don. "Some Considerations on the Archiving of Digital Information", Yale University Library, January 1995; <http://www.ifla.org/documents/libraries/net/waters1.htm> 19/11/2003

Whelan, Karl, "Computers, Obsolescence, and Productivity", Division of Research and Statistics, Federal Reserve Board, February, 2000,
<http://www.federalreserve.gov/pubs/feds/2000/200006/200006pap.pdf> accessed 08/12/2003

Internet Resources of Note

American MicroImaging website, <http://www.amibusiness.com> accessed 16/03/2004

CiteSeer.IST website, Scientific Literature Digital Library, <http://citeseer.ist.psu.edu/cis> accessed 09/12/2003

Global Grid Forum website, <http://www.gridforum.org/> accessed 10/03/2004

Hellenic Ministry of Culture, "Archaeological Museum of Herakleion"
<http://www.culture.gr/2/21/211/21123m/e211wm01.html> accessed 20/03/2004

SETI@home website, <http://setiathome.ssl.berkeley.edu/> accessed 10/03/2004

The Digital Preservation Coalition, <http://www.dpconline.org/>, accessed 27/02/04

The Hyperdictionary.com website, <http://www.hyperdictionary.com/> accessed 18/11/2003

The International Standards Organization (ISO) website <http://www.iso.ch/> accessed 11/03/2004

The Internet Archive Wayback Machine <http://www.archive.org/> accessed 09/04/2004

The Open Process Framework website, <http://www.donald-firesmith.com/> accessed 26/04/2004

The WikiWikiWeb discussion on "Component Definitions",
<http://c2.com/cgi/wiki?ComponentDefinition>. accessed 26/04/2004

The Word Spy website, <http://www.wordspy.com/words/born-digital.asp>, accessed 19/03/2004

Webopedia.com website, <http://networking.webopedia.com/TERM/P/PostScript.html> accessed 11/03/2004

Wikipedia, http://en.wikipedia.org/wiki/Browser_wars accessed 25/02/2004

WordStar Resource website <http://www.wordstar.org/> accessed 11/03/2004

World Wide Web Consortium (W3C) website <http://www.w3.org/> accessed 11/03/2004