3-17-2008

A Multiple Case Study Analysis of Digital Preservation Techniques across Government, Private, and Public Service Organizations

David P. Gough

A MULTIPLE CASE STUDY ANALYSIS OF DIGITAL PRESERVATION TECHNIQUES ACROSS GOVERNMENT, PRIVATE, AND PUBLIC SERVICE ORGANIZATIONS

THESIS

David Gough
Technical Sergeant, USAF

AFIT/GIR/ENV/08-M08

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government.

AFIT/GIR/ENV/08-M08

A MULTIPLE CASE STUDY ANALYSIS OF DIGITAL PRESERVATION TECHNIQUES ACROSS GOVERNMENT, PRIVATE, AND PUBLIC SERVICE ORGANIZATIONS

THESIS

Presented to the Faculty

Department of Systems and Engineering Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Information Resource Management

David P. Gough

Technical Sergeant, USAF

March 2008

AFIT/GIR/ENV/08-M08

A MULTIPLE CASE STUDY ANALYSIS OF DIGITAL PRESERVATION TECHNIQUES ACROSS GOVERNMENT, PRIVATE, AND PUBLIC SERVICE ORGANIZATIONS

David P. Gough

Technical Sergeant, USAF

Approved:

//signed//	17 Mar 2008
Alan R. Heminger, USAF (Chairman)	Date
//signed//	17 Mar 2008
Dennis D. Strouble, USAF (Member)	Date

Abstract

The process of record keeping has evolved through time. As our technology advances, so does our ability to manage information. We have progressed from paper-based records to new digital techniques and formats to store records. However, digital storage is not the "Holy Grail" answer to preservation and storage problems. Digital storage is confounded by multiple problems, also. Some of these problems are, but not limited to, lack of standardization and legal guidance, proprietary formats, and the fragility of the digital medium. This research examines several organizations that are deeply involved in digital preservation and tries to identify common practices and problems across the industry.

Acknowledgments

I would like to thank Dr. Alan Heminger for his support and sponsoring of this effort. Also, I would like to thank Major Jason Turner and Dr. Dennis Strouble for their constant feedback, support, and time. Also, I would like to thank all of the organizations and interviewees that allowed me to work with them for this research effort. Without their support and willingness to participate, none of this would be possible.

I would also like to thank my wife and my children for their love, support, and unceasing patience with my effort to complete this work. Words cannot justify the love and appreciation I have for my family.

Finally, nothing could have been accomplished without the grace of God. His constant presence and encouragement made all the difference in completing this work.

David P. Gough

Table of Contents

Page

List of Figures

Page

List of Tables

Page

A MULTIPLE CASE STUDY ANALYSIS OF DIGITAL PRESERVATION TECHNIQUES ACROSS GOVERNMENT, PRIVATE, AND PUBLIC SERVICE ORGANIZATIONS

Introduction

"*Those who cannot remember the past are condemned to repeat it*."

George Santayana, *The Life of Reason, Volume 1*

Background

Today we are amidst an age of innovation, and technology breakthroughs are commonplace in society. Rarely does the populace approach an invention with amazement any more, and when such an occurrence does happen, the enthusiasm over such an event soon wears off. When the first space shuttle was launched twenty years ago, it captured the attention of the whole world whereas any current launch today barely makes the evening news.

Society's jaded view has some historians alarmed because when no one worries, no one cares. They are beginning to ask "Who is watching our digital history?". Most users think that once a file or a picture is saved to a disk, the problem is taken care of. The information can be accessed whenever the user needs it. However the reality of the situation is that the data might have been saved, but it is by no means safe.

The ability to reach back historically is what we base our collective knowledge store in society. We can trace our steps through history from the past to the present to

form a continuous unbroken chain of events. However, we are in real danger of breaking that chain and losing our history and with it our identity.

The push of technology is the engine that drives the pace of innovation. As a result new technologies are being developed at an incredibly fast rate, and business and industry—the driving forces behind an economy—are reluctant to slow down. To do so would nullify their competitive advantage and rob them of profits. Instead they press forward trying to further distance themselves from their competitors. This relentless push for innovation is focusing on the development of information tools to facilitate their myriad business processes. These new tools make the advancements in technology possible, but on the other hand, sometimes the tools do not mesh well with other older systems. This causes many problems with regards to compatibility and interoperability across computer systems and software platforms in an organization.

An important part of the technology push is the move towards digital storage mediums. The move to digital is essentially inevitable due to the many advantages offered through this medium. Digital media is more portable, more secure, and infinitely reproducible. However, the ability to access information from one type of preservation medium to another has become more difficult as technology advances. One might even conclude that modern technological innovations are hindering our ability to reach back to older storage mediums—gone are the days of the floppy disc and woe to those who still have precious information stored in those places.

Also for the first time in history, society has begun to employ digital preservation strategies to store its historically significant information. From an economic standpoint,

this is perfectly logical. Paper copies require large physical storage facilities, file cabinets, storage lockers, and boxes and folders along with environmental controls to temper the deterioration of that medium. Compare this type of undertaking to electronic storage where capacity is vastly increased, cost dramatically reduced, and data sharing is augmented. It seems like going digital is an easy enough conclusion to reach. Digital benefits include: simultaneous usage of information, ability to manipulate information, and more options for viewing, printing and storing information (Ireland, 1998). However, a problem arises when you consider how people will access the digitally stored information that was stored on devices or software that are now obsolete. Getting at digital information is not as easy as accessing paper based information. Almost anyone can pick up a piece of paper and read the information recorded on it all by themselves. The information can be plainly seen and interpreted. However, digital information is not easily seen, and to interpret digitally stored information requires electronic devices and appropriate software. This creates problems as systems, devices, and software become obsolete.

Digital data uses binary data strings (0s and 1s) to represent data. A "bit" is the smallest unit of information that can be stored in a computer, and consists of either a 1 or 0—the "on/off" state. All computations performed by the computer are expressed in bits. A "byte" is a collection of 8 bits. Bytes are convenient because when converted to computer code, they can represent 256 characters such as numbers or letters. Common aggregations for bytes come in multiples of 1000, such as kilobytes, megabyte, gigabyte, and so on. These aggregations are displayed in Table 1.

Amount	Size
Bit (b)	1 or 0
Byte (B)	8 bits
Kilobyte (KB)	1,000 bytes
Megabyte (MB)	1,000 KB
Gigabyte (GB)	1,000 MB
Terabytes (TB)	1,000 GB
Petabyte (PB)	1,000 TB
Exabyte (EB)	1,000 PB
Zetabyte (ZB)	1,000 EB

Table 1: Data grouping levels

The digital format that the computer industry uses today is based on Boolean logic. If you look back at the history of computer technology, you find that all computers are designed around Boolean gates—the NOT, AND, and OR operations. The technologies used to implement those gates, however, have changed dramatically over the years. The very first electronic gates were created using relays. These gates were slow and bulky and were eventually replaced by vacuum tubes. Tubes were much faster but they were just as cumbersome, and they were also plagued by the problem that tubes burn out—just like light bulbs. Once transistors were perfected, however, computers started using gates made from discrete transistors. Boolean logic in the form of simple gates is very straightforward. From simple gates you can create more complicated functions, like

addition. From those three facts, you have the heart of the digital revolution, and you understand, fundamentally, how computers work.

Digital information is stored electronically, and to decipher digital information requires some type of decoder, but all too often you need the exact type of decoder. Consequently there is growing concern among IT professionals that society is approaching a point when we will not be able to reference data and information from the past and, tragically, it will be lost to us forever.

To illustrate this point, let us consider a number of recent technologies. The 8-track tape was a fairly popular commercial medium that primarily was used by the music industry to dispense their product to the public. The popularity of the 8-track and vinyl records reached their zenith in the 1970s. Today those media are no longer used to record music, and 8-track players are difficult to find today. The life of the 8-track was cut short by the invention of the cassette tape—a smaller, more portable unit that surpassed the 8-track tape in quality and affordability. Further, the cassette tape, while immensely popular in the 1980s, has all but vanished from store shelves. The compact disc replaced it as the new medium of choice for the consumer. But even today, CDs are waning in popularity due to the rise of the DVD and MP3 devices.

Although the plight of CDs and cassette tapes might seem trivial, the preservation problem is spreading further than society realizes. Technically advanced organizations such as NASA have felt the effects. Up to 20% of the information carefully collected on Jet Propulsion Laboratory computers during NASA's 1976 Viking mission to Mars has been lost (Stepanek, 1998, 1998). The lost data was trapped

on decaying digital magnetic tape, forcing NASA to call mission specialists out of retirement to help the agency reconstruct key data elements (Stepanek, 1998).

NASA was fortunate enough to recover most of their mission data; however, there are other examples of organizations that were not so lucky. Some POW and MIA records and casualty counts from the Vietnam War, stored on Defense Dept. computers, can no longer be read (Stepanek, 1998). And at Pennsylvania State University, all but 14 of some 3,000 computer files containing student records and school history are no longer accessible because of missing or outmoded software (Stepanek, 1998).

The problem of losing information from the past is not only tied to the data, but the technology as well. In September 2007, a documentary film, "In the Shadow of the Moon", was released in the United States. The film documents the Apollo Space Program from conception to close. In one question and answer session before a particular premier, the director, David Sington, was asked a question in light of talks about NASA preparing a mission to return to the moon. How does the state of technology used during the Apollo program measure up against technology today?

Q: How basically does the technology compare to today?

DS: Well we have more computing than on the Lunar Excursion Module (LEM). The LEM computer got overloaded. And I think there is 38 kilobits of addressable memory which is a tiny, tiny

> fraction of what is in your mobile phone. In some ways its interesting talking to the NASA people about going back to it [the moon]. They don't know how they first did it. These techniques and the technology have been lost. Rediscovered and reinvented. Its a little bit like that wonderful blue you get in medieval stained glass. Nobody knows how to do it anymore.

Perhaps "tragic" is an inappropriate term to describe the 8-track's demise, but consider Martin Luther King's *I Have a Dream* speech made during the Civil Rights Movement in this country. What if that pivotal piece of history was lost?

Another problem to be addressed is the enormity of such a task: the undertaking of preserving the vast quantities of documents. The National Archives and Records Administration (NARA) is a federal agency that has considerable resources, both financial and legal, at its disposal. NARA was created as a repository for government documents and other historically significant materials and is in a unique position to address the digital conversion problem. The very mission of NARA makes it a pivotal stakeholder to formulate a solution to the digital preservation problem.

As defined in Section 3301 of the US Code (and similarly for Presidential records in Sections 2111 note and 2201), the types of records NARA is responsible for are:

> *all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that*

> *agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them.*

"NARA faces increasingly enormous quantities of records" (Carlin, 1998). As if sheer volume was not enough of a problem, NARA is also receiving "an increasingly diverse load of digital information" created using a wide variety of software and stored in a "bewildering variety of media" (Smith, 1998). This predisposes information to the threat of being permanently lost, even if it is under NARA's watchful eye. Also it appears that the amount of data that society creates is growing exponentially (Carlin, 1998). The sheer magnitude of new data that are being added to the already large store of digital information exacerbates the problem of managing it all. The Archivist of the United States put it eloquently when he said, "It will be worse than sad if the marvelous technologies that are giving us a new information age outrun our ability to keep a record of it" (Carlin, 1998).

To give you an idea of where all these exabytes come from, just consider the number of devices or subscribers in the world that can create or capture information. Table 2 displays a partial list of such devices.

Type of Device	Number in Millions in 2006
Camera Phones	600
PCs	900
Audio Players	550
Mobile Subscribers	1600
LCD/Plasma TVs	70
Digital Cameras	400

Table 2: Digital Devices

By 2010, this installed base of devices and subscribers will be 50% larger, devices will be cheaper, and resolutions higher. All of these devices are creating more and more digital bits. Figure 1 shows the expected growth of the digital landscape.

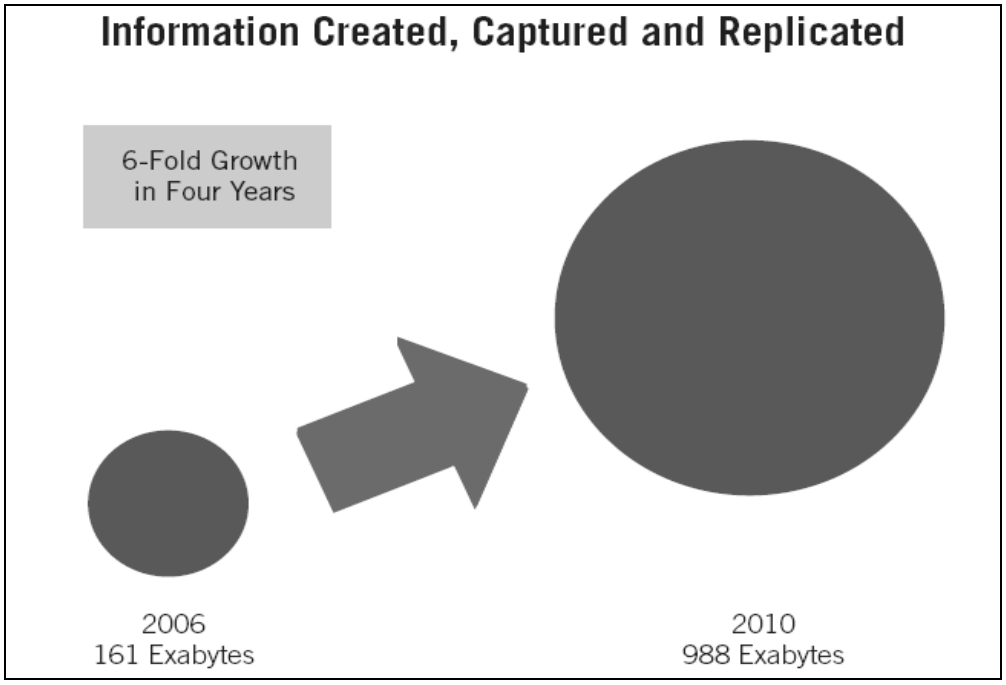


Figure 1: Digital information growth (IDC, 2007)

Protecting digital information "requires the preservation of the knowledge and technology necessary to access" the information (Robertson and Heminger, 1996). One method proposed to solve this digital dilemma was offered by Steve Robertson and Alan Heminger. Robertson (1996) developed a digital preservation model called the Digital Rosetta Stone (DRS). "A DRS is composed of three major processes that are necessary to preserve and access our digital history—knowledge preservation, data recovery, and document reconstruction" (Robertson and Heminger, 1996). The key process is knowledge preservation. It provides an essential basis for data recovery and makes document reconstruction possible. In this process information about media storage techniques and file formats is gathered and recorded in a metaknowledge archive (MKA)

(Robertson and Heminger, 1996). When required, pertinent information concerning unique formats is gathered from the MKA. This information, in turn, is used to recreate digital documents.

Problem To Be Addressed By This Research

Several concerns arise due to the migration towards digital preservation in this country and the world. Due to the growing quantities of digitized information, organizations are turning towards digital preservation methods to manage their documents. Potential problems exist due to ever changing program formats and the unstable digital storage mediums. In light of these concerns, the DRS model was developed to help solve this preservation problem. Just as the original Rosetta Stone was used to unlock the mysteries of stored written information, The Digital Rosetta Stone model was proposed to recover the digital bitstream from an obsolete medium and interpret that bitstream so the information can be properly displayed (Robertson and Heminger, 1996).

However, it has been noted from previous research in this area that "outside the library and archival communities there has been almost no recognition of the importance of long-term preservation (here called archiving) of electronic information" (Ireland, 1998). "Within the library community there are several important digital library projects addressing archiving as a critical need, but they are uncoordinated and to date publicly almost mute" (Ireland, 1998).

To answer the call, several private institutions have stepped into the preservation

arena. These organizations are actively developing digital preservation strategies to reduce operating costs and streamline business functionality. Currently, there are several types of methods to achieve this process, and because of the economic ramifications, an investigation is warranted to identify strengths and weaknesses of various storage strategies.

This study focuses on three disparate organizations that were selected to represent different parts of this country's society and economy. The three sectors to be analyzed in this study are private, for-profit organizations; governmental organizations, and public service organizations.

Research Question

What is the current state of digital preservation efforts across the three major sectors of U.S. society?

Literature Review

Down through the ages civilizations have recorded history albeit in myriad, and sometimes obscure, ways. The history has been written on paper and papyrus, inscribed in tablets and pottery. These have been handed down from generation to generation entrusted from father to son and mother to daughter. But sometimes through accident or negligence the history is lost. It leaves a hole in time that many feel compelled to fill. It is a chain that is missing a link. Have any of us ever wondered about the lost city of Atlantis? There are a scant few threads of evidence to support its existence, but other than that we do not know anything about it really. Or was there really a King Author of England? Again there are obscure clues and scraps of information but nothing to confirm his existence. It is surprising that these more interesting episodes in the world's history do not have more evidence associated with them. Undoubtedly the knowledge existed at one time, but has vanished from the collective knowledge of history. Why? The simple answer is that no one was there to protect it and safeguard it. Society must take care to preserve itself. Some civilizations did take those valuable steps leaving treasures to be found by later generations.

In a 2005 article, Ken Quick and Mike Maxwell describe the Rosetta Stone and the need to have the equivalent to capture archived encoded data.

> *A black basalt slab with strange inscriptions on it, the Rosetta Stone was unearthed in July 1799 by Napoleon's army in Rosetta (Rashid), Egypt, and kept as a souvenir by one of the troops through generations of the soldier's family. Eventually, it ended up in a flea market and was spotted*

by an Oxford professor of Egyptology on vacation in France who recognized hieroglyphics on a portion of the stone. It was quickly discovered that the Rosetta Stone contained the same text in 3 languages and one was Greek. It was the key to decoding the there-to-fore undecipherable hieroglyphics, the archived writing of the ancient Egyptians. (Quick and Maxwell, 2005)

According to Cornell University Library's somewhat chilling "Obsolete and Endangered Media Chamber of Horrors," no fewer than thirty-two distinct media formats for backing up digital information have emerged since the advent of modern computing. This includes once seemingly impeachable technologies such as the 5 1/4 inch floppy, Sony's line of WORM disks, Syquest cartridges, and IBM's half-inch tapes (superseded by quarter-inch tapes). Even formats that seem to live forever, such as the 3 1/2 inch "floppy" introduced with the first Macintosh computer twenty years ago, have declined in popularity and will soon join the ranks with its preceding formats in the dustbin of history. Dell Computer, the world's largest computer company, recently dropped the 3 1/2 inch floppy from its line of desktop and laptop computers (Cohen and Rosenzweig, 2005).

Digital vs. Paper

As a simple method of preserving a digital document, why not print it out and save it as hard copy? Rothenberg rightly points out that this is not a complete solution

since some multimedia documents cannot be properly printed out to capture all of its properties and any interactivity a document possesses will be lost (Granger, 2000). Also, "Many types of digital objects do not have print equivalents and cannot be preserved in non-digital formats" (Hedstrom, 2003).

"Probably the single most important issue posed by electronic records to archives and archivists is how to ensure access over time to records in electronic form when the information technologies themselves are subject to rapid and sometimes revolutionary change" (Dollar, 2002). Electronic resources are profoundly unstable, far more unstable than paper records. Crucial information has been gathered from letters and photographs weathered from exposure to decades of sunlight, from hieroglyphs worn away by centuries of wind-blown sand, and even from papyri partially eaten by insects. In contrast, a stray static charge or erratic magnetic field can wreak havoc on the media used to store "digital objects". It is possible that out of the millions of bits that comprise a digital file, the corruption of only a few may make the file unreadable and unusable. With some exceptions, digital formats tend to require an exceedingly high degree of integrity in order to function properly. This creates an intriguing paradox. A digital object's perfection is also its imperfection. These files are encoded in such a precise fashion that unlimited perfect copies may be generated, but any loss of their perfection may leave them useless and can mean disaster (Cohen and Rosenzweig, 2005).

Paper documents also utilize logical structures and physical relationships. The logical structure of a document includes words, sentences and paragraphs. Font type,

margins, spaces, headings, etc., define the physical relationships between portions of a document.

In contrast, the reader of a digital document is unable to take advantage of the same flexibility offered to readers of paper documents. "The arrangement of the electronic signals that form [an electronic document] seldom bears any relationship to the image displayed on a monitor" (Dollar, 2002). "The allocation of storage space for electronic information is a function of computer operating systems that store bits and pieces of data wherever space is available" (Dollar, 2002). Figures 2 and 3 show various types of electronic storage currently used today.

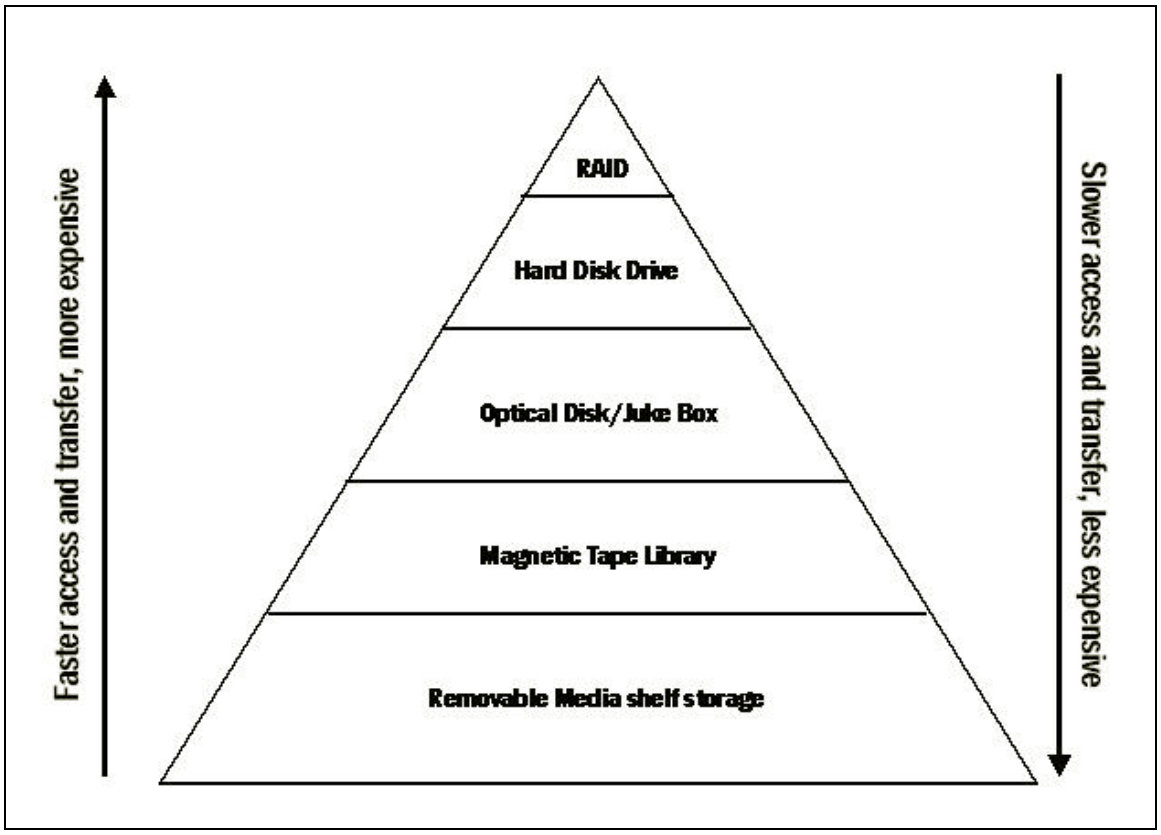


Figure 2: Types of Digital Storage

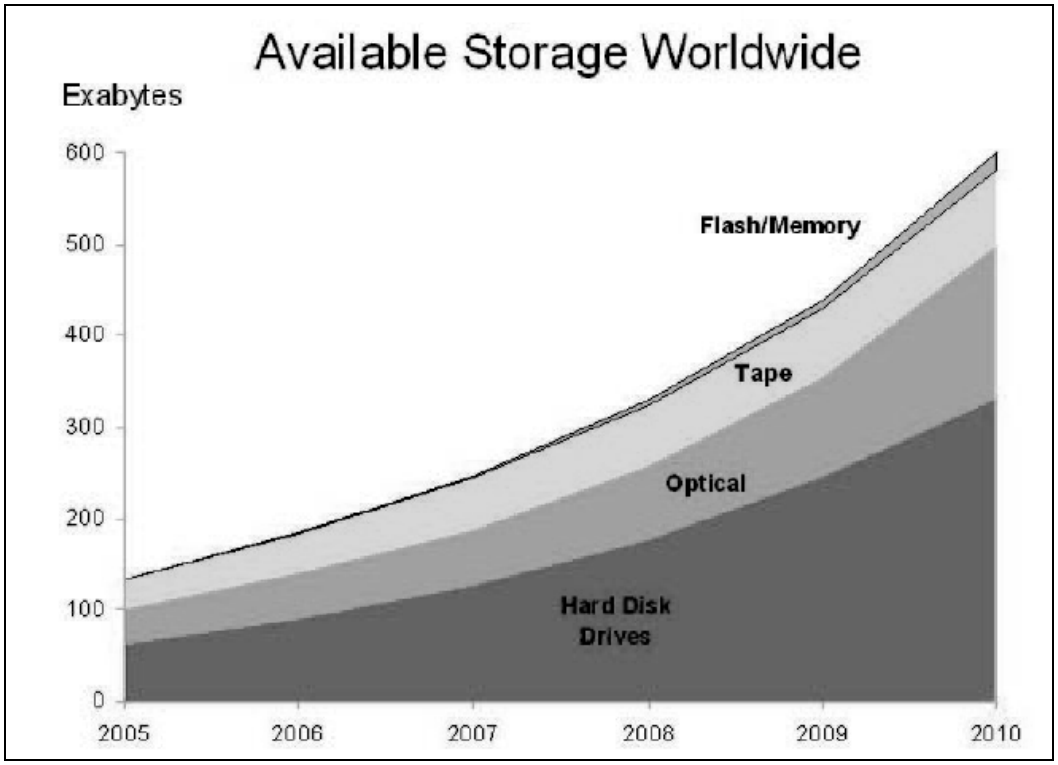


Figure 3: Available Storage (Gantz, 2007)

There are "trade-offs between what is desirable from the standpoint of functionality, dependability, and cost and what is possible and affordable with current technologies" (Hedstrom, 2003). Due to decades of obsolete computers and drastic changes in formats over the years, digital information is at enormous risk of being lost. However, there are methods of recovering information from obsolete software and research is being done to find new techniques of preserving the world's data.

A number of people have proposed strategies to address the threat to our stored digital information (Robertson and Heminger, 1996). However, there is no agreement on which method is the best way to proceed (Kochtanek and Hein, 1999). The machines

used to create records in the late 20th century have long been replaced by technologies that are faster, better, and cheaper. As demonstrated by Figure 4 below, Moore's Law shows that as technology doubles every 18 months. This technology curve adds to the obfuscation of current preservation formats. However, the issue of thousands of data formats is only part of the challenge (http://www.archives.gov/records-mgmt). A standardized approach must be agreed upon for the preservation effort to be successful.

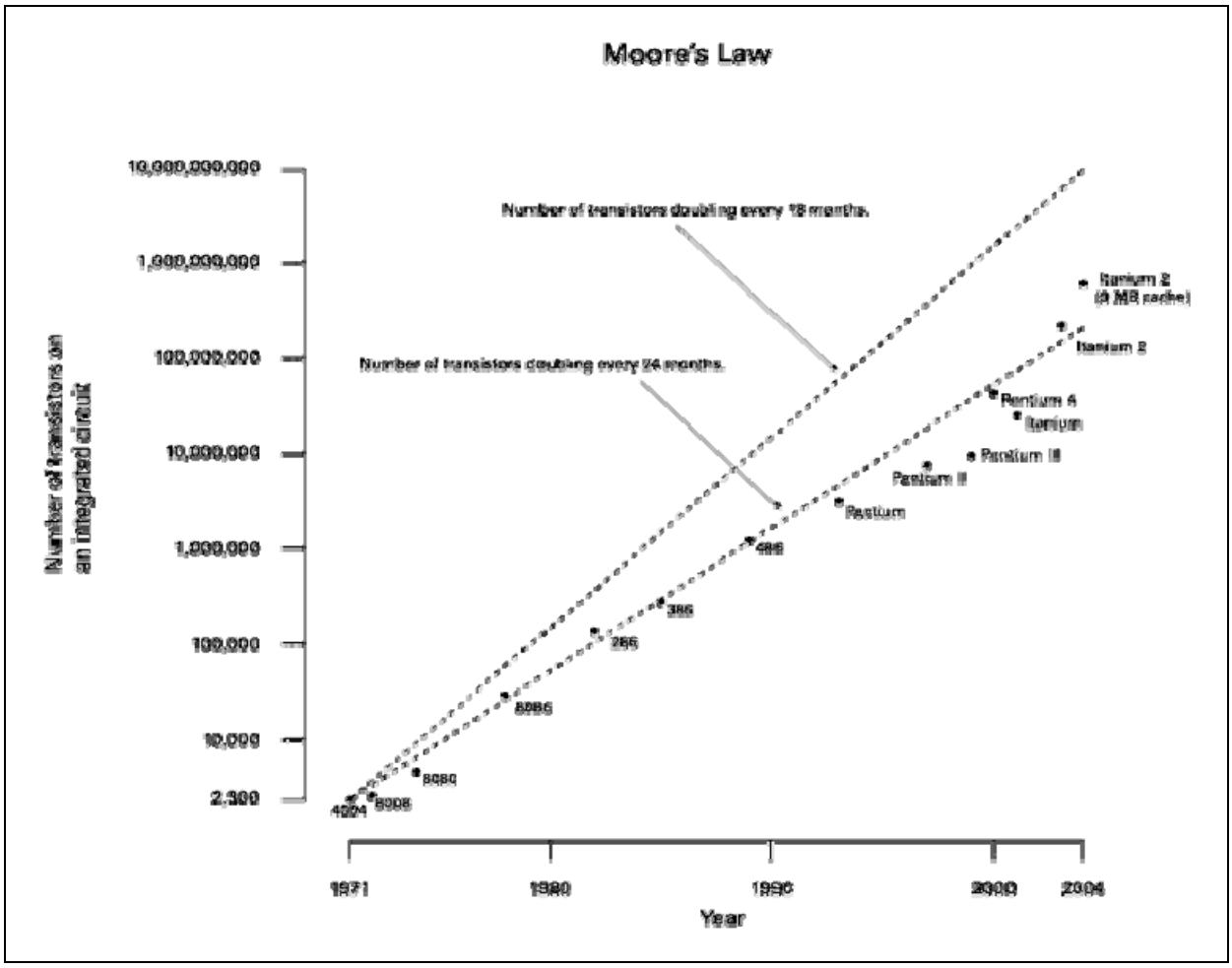


Figure 4: Moore's Law

The efforts of NARA and others to date have been labor intensive and expensive. Currently, many organizations are researching the best options to store digital information so that it is not lost in the shuffle in the constant updating of hardware and software. The rate of technological obsolescence is such that records created and accessed even two years ago may now be unreadable. Unless this challenge is confronted and surmounted, there will be no National Archives for the digital era (http://www.archives.gov/era/about/).

By 2010 the installed base of digital devices and subscribers creating and collecting information will be 50% larger (Gantz, 2007). Devices will be cheaper, and resolutions higher, all creating more and more digital bits (Gantz, 2007). How much of the information that is captured, created, or replicated also is stored is another matter. As part of the research for this project, information was gathered to predict how much storage will be available to store all this information, should we choose to. Figure 5 shows the contrast between the amount of digital information that is being created and the storage resources available to maintain it.

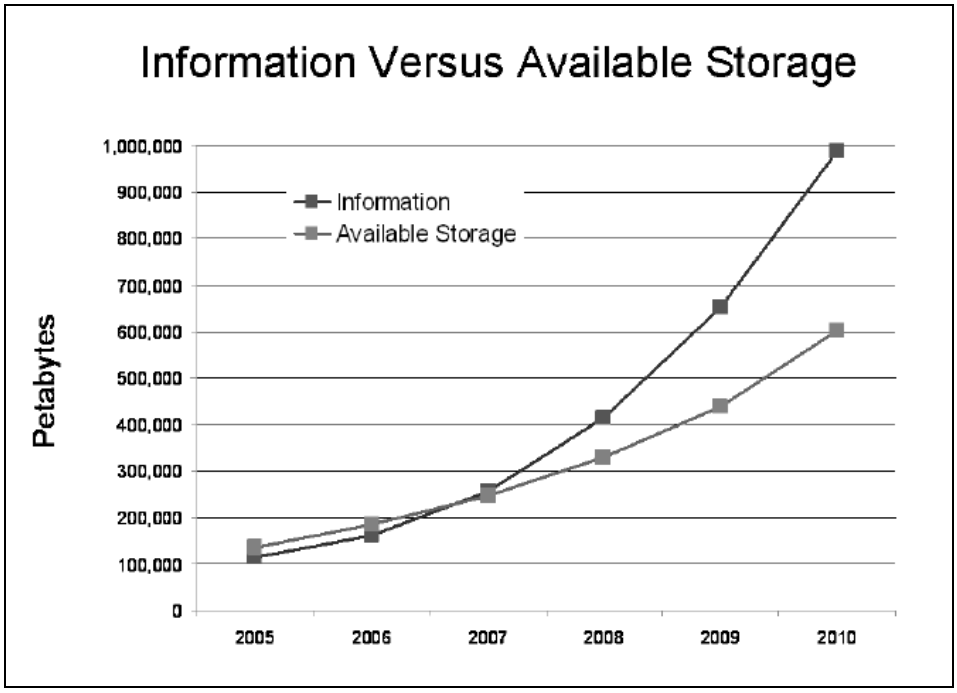


Figure 5: Information vs. Storage (Gantz, 2007)

NARA's Strategic Response was the development of the Electronic Records Archives (ERA). The ERA project was started in 1998. The first three years were spent researching the problems with digital preservation and to see what might be possible. The research activities have all been collaborative with other Federal Government agencies, state governments, computer scientists, other national archives, academia, and private industry. One of these collaborations was in the development of the Open Archival Information Standard (OAIS). ERA will be an OAIS because the OAIS standard addresses any kind of information kept for any length of time (http://www.archives.gov/era/about/).

ERA will provide NARA with the capability to authentically preserve and provide access to any kind of electronic record free from dependence on any specific hardware or

software. The research and development have provided NARA with the ability to transfer, preserve, manage, and provide sustained access to all types of electronic records (http://www.archives.gov/era). The ERA System will ensure that anyone, at any time, has access to find and use the electronic records that NARA preserves (http://www.archives.gov/era).

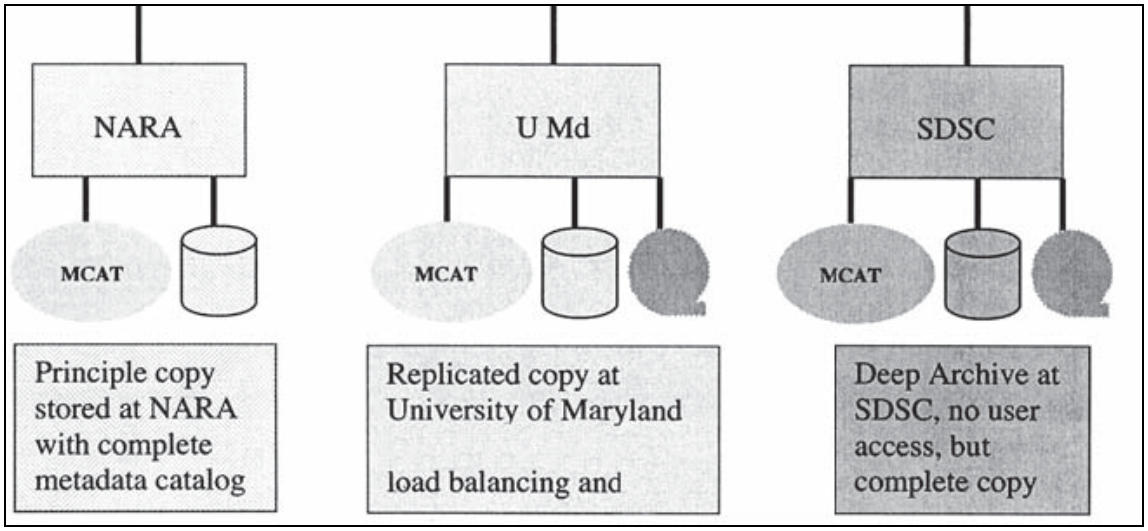
Examples of ERA's many benefits are to:

- Enable intelligence agencies to better store and share vital historical security information;
- Access geographic information systems and virtual reality models used in some patent applications;
- Ensure the electronic records documenting drugs and related research data are accessible over time to the Food and Drug Administration;
- Permit the Federal Aviation Administration full access to electronic aircraft safety records;
- Facilitate the repair of the Department of the Navy's damaged ships, whose design records are only available as electronic computer aided design (CAD) files;
- Share drug application information between pharmaceutical companies and the Drug Enforcement Administration.

The NARA ERA implements a preservation environment that is designed to mitigate against all of types of risk. The system is depicted in figure 6. The ERA preservation environment is achieved through the integration of three independent data grids. Each grid manages its own preservation metadata in a separate metadata catalog. Consistency constraints are implemented between the grids to control which digital entities may be replicated between the data grids, how the preservation metadata is

synchronized between the data grids, and how user identifiers are replicated between the data grids.

Figure 6: NARA's Prototype ERA Configuration



• *Multiple copies are kept at University of Maryland (U Md). U Md uses a High Performance Storage System (HPSS) to replicate files that are provided for public access on a commodity-based disk storage system. This mitigates against media loss.*

• *U Md replicates data onto a commodity disk system at NARA. This protects against operational error at U Md and protects against simultaneous loss of the two copies at U Md. The U Md and NARA metadata catalogs are implemented in different database technologies (Informix and Oracle) to protect against systemic vendor product failure.*

• *A copy is replicated into a deep archive at SDSC. This protects against natural disaster (tornados), and also provides a copy that has restricted access to protect against malicious users.*

The Library of Congress (LOC) has the responsibility to collect and preserve our cultural and intellectual artifacts. In December 2000, Congress authorized LOC to develop and execute a National Digital Information Infrastructure and Preservation Program (NDIIPP). This included collaborating with other Federal and non-Federal entities to collect and provide access to digital materials and developing a strategy for the policies and technological infrastructure needed to insure long-term preservation (Schloman, 2003). On February 14, 2003, the Librarian of Congress announced that Congress had approved the NDIIPP plan. Already developed is a prototype system to collect and preserve materials from the Web, "Minerva" (Mapping the Internet: the Electronic Resources Virtual Archive) is concerned with Web materials that have been made publicly available without restriction (Schloman, 2003). News services, such as CNN, are sites that are publicly available, but request that Internet robots exclude them from their information harvesting sweeps. Figure 7 illustrates the design of the NDIIPP system.

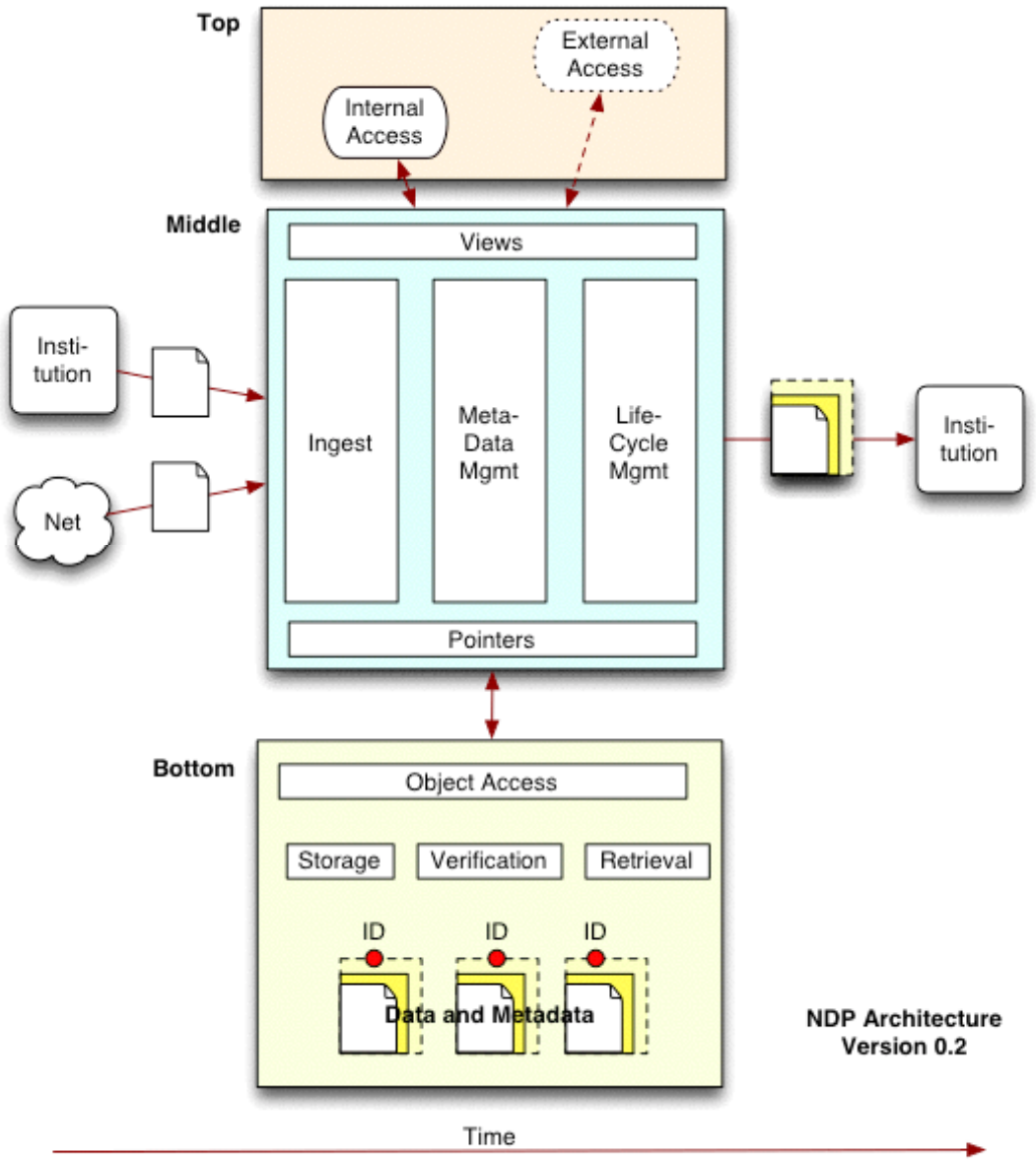


Figure 7: The NDIIPP

The National Institute of Standard and Technology (NIST) is another organizations concerned with the growing problem of possible data loss, degradation, or alteration. Currently, the amount of information produced, lost, and corrupted in organizations is a massive amount of data. Engineering drawings with computer-aided

design (CAD) and other computer generated data that should be maintained and shared with future engineers is losing terabytes of information because of a lack of digital storage standards (Jackson, 2006).

Federal agencies, also, are creating an exponentially increasing volume of diverse and complex digital records. As electronic record formats become more and more sophisticated, they grow increasingly difficult to manage. These new technologies have the capability of evolving rapidly, making older technologies and formats obsolete. The more complex the materials, the more difficult they are to preserve, and no current system exists to preserve these digital records over time. Consequently, the possibility to lose many electronic records exists (Cacas, 2005).

In the midst of these technology fluctuations, new legislation is being generated by Congress requiring companies to retain e-mail and other types of electronic records for reporting purposes. Motivated by the rise of large corporate scandals, Congress passed the Sarbanes-Oxley Act which profoundly changed the way corporate America does business. Due to their desire to respond to this economic crisis, Congress also redefined the laws surrounding stock market securities. This resulted in more statutory amendment revisions since the original 1933 and 1934 securities laws. The changes addressed perceived shortcomings in the law's ability to properly handle abuses such as false financial reporting, auditors not following sound accounting practices, and the destruction of evidence. The real-world examples of these abuses were corporate scandals in companies such as:

- Enron, which used off-the-books partnerships to inflate earnings and conceal debt;
- Arthur Anderson, which failed to stand up to Enron's aggressive accounting techniques and later destroyed documents relevant to the investigation; and
- WorldCom, which overstated its revenue and understated its losses.

The Sarbanes-Oxley Act addresses the threat of fraud in the finance departments of public companies by requiring companies to establishing suitable "internal controls" for gathering, processing, and reporting financial information. The ultimate goal of the act was to ensure companies' finances for the benefit of investors. However, the far-reaching effect of the law has created profound ripple effects in the IT departments of corporations. Companies have had to implement preservation strategies based on the legal requirements placed on them by this act.

As demonstrated in Figure 8, IT expenditures are increasing albeit slowly throughout the economy.

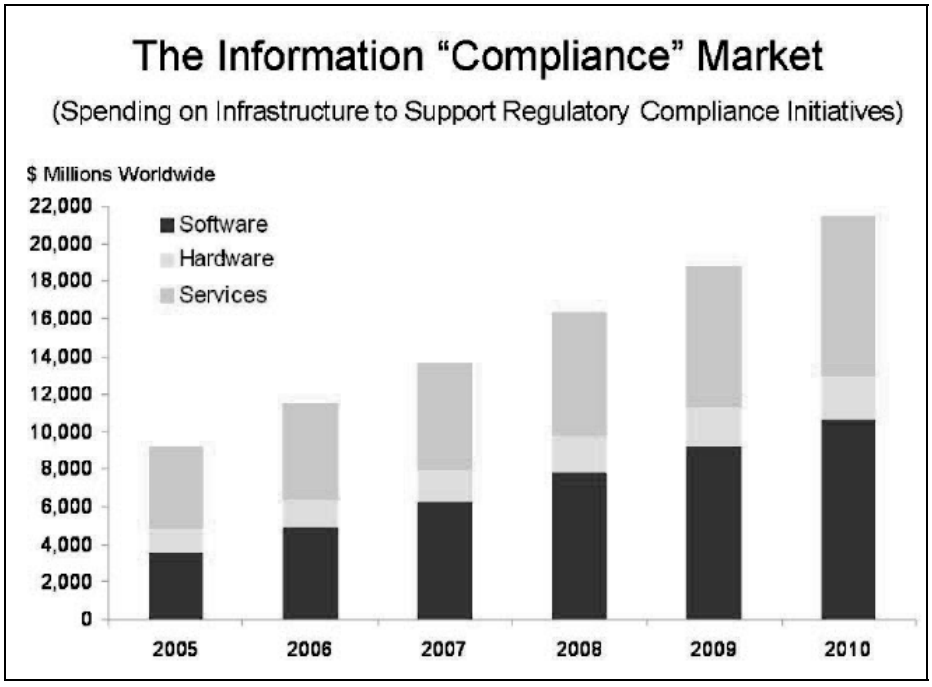


Figure 8: Dollars spent on IT compliance (Gantz, 2007)

In a recent survey by the Association for Information and Image Management (AIIM) indicates storage consumers are badly informed when it comes to archiving. Of 1,000 organizations surveyed by AIIM, 45.9% considered email archiving the responsibility of individual employees; 25.5% considered it part of an overall information management strategy and 8.4% saw it as a stand-alone application.

AIIM's report on the survey says most organizations consider archiving as a collection of massive backup files. In conjunction with this finding, a July report in Byte and Switch Insider also addressed the fact that organizations confuse archiving and backup.

According to that report, backup involves making point-in-time copies of data to protect against hardware failures or catastrophic data loss and includes operating systems as well as applications. Archiving, the report maintains, requires fast file-level access to data and should involve search and retrieval software and indexed repositories. Whereas backup volumes are usually kept for days before being replaced by new volumes, archived files can be kept for decades.

In other words, archiving applies to data that is likely to be needed again and logically placing it where it can be easily found and retrieved. Backup creates a mirrored copy of data; a safety resource to be used in the event that something happens to the original data. Backup is for recovery while archiving for preservation and retrieval. Backup is short term and data can be periodically overwritten; archiving is long term and unalterable and thus used for compliance. Table 3 shows the differences between "backup" and "archive".

Use Case	Back-up	Archive
Subpoena for information contained in email	Restore all backup files for requested time period • Affects production system • Degrades performance • Is it possible? • Does back-up scheme contain provisions to store information for the legally required time frame?	Simple query of system
Extract information for delivery to requester	Locate required emails and restore as above • Establish "legal" mailbox • Extract to industry standard format	Establish case file and output
Continue to retain information that is required for active litigation	Impossible within standard procedures of backup and recovery plans	Mark required emails with "litigation hold"
Eliminate emails that no longer legally need to be retained	Impossible within standard procedures of backup and recovery plans	Classify emails in a records retention plan and automatically delete
Prevent emails containing sensitive information and intellectual property from being easily removed from the system	Impossible while still providing regular business access to older emails	Easily configured, using archive for long-term retention of older emails
Ensure cost effective satisfactory system performance	Difficult and costly when long-term history is stored within the email system and archive files are located on individual desktops	Easily configured by offloading older emails to the archive system

Table 3: Use Case Scenarios – Email Back-up versus Email Archive, "2006 E-mail Management: An Oxymoron?" AIIM Study, 10/11/2006

APPROACHES/TECHNIQUES

There have been significant amounts of research invested in solving the digital preservation problem. Some of the approaches are described here.

Micrographics

This solution seeks to solve the problem of requiring vast amounts of paper resources and reducing the storage facilities that would be needed to care for paper-based documents. This strategy is similar to paper storage except the medium is plastic and the information is miniaturized. Some of the major advantages with this process are that it is already used as an archival medium with well-documented standards. It is easy to read the medium, and it can store a high resolution of detail. Microfilm is the Rosetta-Stone equivalent media suggested by Quick and Maxwell because it can capture "human readable standard text images with illustrations, tables, formulae, and example computer source code…along with coded binary data". Microfilm lasts 500+ years, has a low dependency on hardware, and provides compact storage.

Microfilm can be used to save information, but it must be able to be referenced to be useful. A simple-text database can be imported into any current or future search engine to accomplish the task of subject retrieval in a microfilm archive. However, the database would need an exhaustive list of the file locations in any collection (Quick and Maxwell, 2005). This might prove to be an arduous task.

Also, another shortcoming of this medium that has only developed recently is the viability of the medium. According to the University of California at Berkeley, ink-on-

paper content represented an incredibly miniscule 0.01 percent of the world's information produced in 2003, with digital resources taking up over 90 percent of the non-printed majority (Cohen and Rosenzweig, 2005). It is not prudent to continue to carry legacy technology that's usefulness is quickly eroding.

Other disadvantages arise, however. This medium has to be physically handled to access the information. Consequently, it suffers from the same possible damages that occur with paper-based document in storage or use.

Nickel Slugs

Because paper and plastic tend to deteriorate when handled and are subject to limited environmental conditions, some have suggested engraving the digital information on nickel slugs would be a viable solution (Rothenberg, 1998). These metal slugs have a great amount of resistance to deterioration lasting for thousands of years. However, transferring information to non-electronic media makes it far more difficult to access the information.

The *New York Times* felt compelled to use an analog solution for their millennium time capsule, created in 1998-1999. The *Times* bought a special kind of disk, a HD-Rosetta, pioneered at Los Alamos National Laboratory to withstand nuclear war. The disk, holding materials deemed worthy for thousand-year preservation by the editors of the *Times* magazine, was created by using an ion beam to carve letters and figures into a highly pure form of nickel (Cohen and Rosenzweig, 2005). Again, there is the problem of accessing the data stored on the disk. Will the hardware be available to read the

information in the future? There are no guarantees. The ageing hardware concern brings us to the next proposed preservation solution.

Technology Museums

The disadvantage of digital storage results from the obsolescence of the hardware and software technology used to create and subsequently, to read the information (Ireland, 1998). To combat this problem, another approach was offered to preserve our posterity. This method seeks to store every generation of technology, keep the machines in working order, and run them with skilled operators. This is referred to as a technology museum. This approach would extend the longevity of computer systems and their original software to keep documents readable (Rothenberg, 1998). Software used to store information often is written for a specific hardware platform (Ireland, 1998).

However, disadvantage of a technology museum is that "the hardware and software for digital media change so rapidly that it would be impossible to keep an up-to-date museum" (Pace, 2000). Rothenberg (1998) argues that "even if obsolete computers were stored carefully, maintained religiously, and never used, the natural aging and decomposing processes would eventually render them inoperative; using them routinely to access obsolete digital documents would undoubtedly accelerate their demise."

The computing community has made strides toward extending the "life-expectancy" of software and making it more ubiquitous. "The programming language XML is a good example resulting from such efforts. Extensible Markup Language (XML) is a popular programming language whose primary purpose is to facilitate the

sharing of structured data across different information systems, particularly via the Internet" (Bray, et. all). XML will be discussed more thoroughly later in this paper.

Emulators

Emulation refers to the ability of a computer program or electronic device to imitate another program or device. In a theoretical sense, the Church-Turing thesis shows that any computational environment can be emulated by any other machine. However, in practice, this task can be quite difficult. When technical specifications of the program and the behavior of the system have not been adequately documented, emulation decisions may not be accurate, and consequently an exact duplication of a digital object may not be possible.

These type of problems spill over into another problem area, legal ramifications, regarding emulation and digital preservation in general. An important question to be answered is if digital objects and records cannot be perfectly recreated, are they sufficient? Are they "good enough"? It is important to understand that what archivists refer to as a record is something much stronger than what many of us non-archivists understand (in our naivety probably thinking of a database record as a RECORD) (Granger, 2000). In one of his papers David Bearman says,

> *"…most information created and managed in information systems is not a record because it lacks the properties of evidence. Information captured in the process of communication will only be evidence if the content,*

structure and context information required to satisfy the functional requirements for record keeping is captured, maintained and usable."

Another possible legal pitfall concerns Intellectual Property Rights (IPR) issues that could involve emulating either operating systems or their applications (Granger, 2000). As computers and global computer networks continued to advance and emulator developers grew more skilled in their work, the length of time between the commercial release of a system console and its successful emulation began to shrink. Many companies saw significant amounts of emulated applications of their products, even though the item in question was still very much in production. This has led to a more concerted effort by IT manufacturers to crack down on unofficial emulation.

"In an archive, it may be necessary to handle some emulations, but this can only be tenable in the short term, while both the emulated and the host emulator are current in technology terms. Obsolescence for the host environment will bring double jeopardy for the emulated environment." (Bennett, 1997) Once the original application has passed, can the emulation be far behind? How can it be effectively maintained once the data it borrows from is lost? "Archiving of an emulation and its dependants should be considered only for the near term, and in the advent of destructive forces" (Bennett, 1997).

Migration

"Migration is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation" (Waters and Garrett, 1998). Although migration is an accepted strategy for alleviating some of the effects of technological obsolescence on long-term storage and retrieval of digital information, it is a difficult and very costly approach (Waters and Garrett, 1998). Although migration can be incredibly expensive, it is a known and effective process (Cohen and Rosenzweig, 2005). However, "it is conceivable that no existing organization can afford the financial, physical, and human resources necessary to carry out such a tremendous task" (Robertson and Heminger, 1996). "The volume of holdings and the need to continuously refresh, duplicate, or migrate data to new formats are likely to place heavy economic burdens" on any organizations (Butler, 1997).

Archivists who have studied the problem of constant technological change realized some time ago that the ultimate solution to digital preservation will come less from specific hardware and software than from methods and procedures related to the continual stewardship of these resources (Hedstrom, 2003). That does not mean that all technologies, file formats, and media are created equal. "Only the strong survive" is an axiom that should also be applied in information technology. But sticking to fundamentally sound operating principles in the construction and storage of the digital materials to be preserved is more important than searching for the elusive digital equivalent of acid-free paper (Cohen and Rosenzweig, 2005).

Standards-Based Approaches

Several strategies regarding a standards-based solution have been proposed and developed by different organizations. Commercial entities recognize the value of these methods. Developing and implementing data format standards may extend the time between migration iterations, and thus organizations may be able to conserve scarce resources (time and money) by advocating for and adopting these standards (Ireland, 1998). The major component of each of these strategies is making access to data easier by only using one or a few methods. Technology experts began to understand that preservation of digital information is "much more dependent on the life expectancy of the access system" than the longevity of the information media (Conway, 1994).

Extensible Markup Language

What is metadata and where does it come from? Could you imagine a person sitting at a desk from 9-5 inputting this metadata in an effort to make e-mail retrievable in the future? No, the number of hours needed to conduct this type of business would not be efficient. Automation of this process is accomplished with Extensible Markup Language (XML). XML is the automation tool for metadata. It can be described as a "self-describing" documentation process (http://www.archives.gov/era).

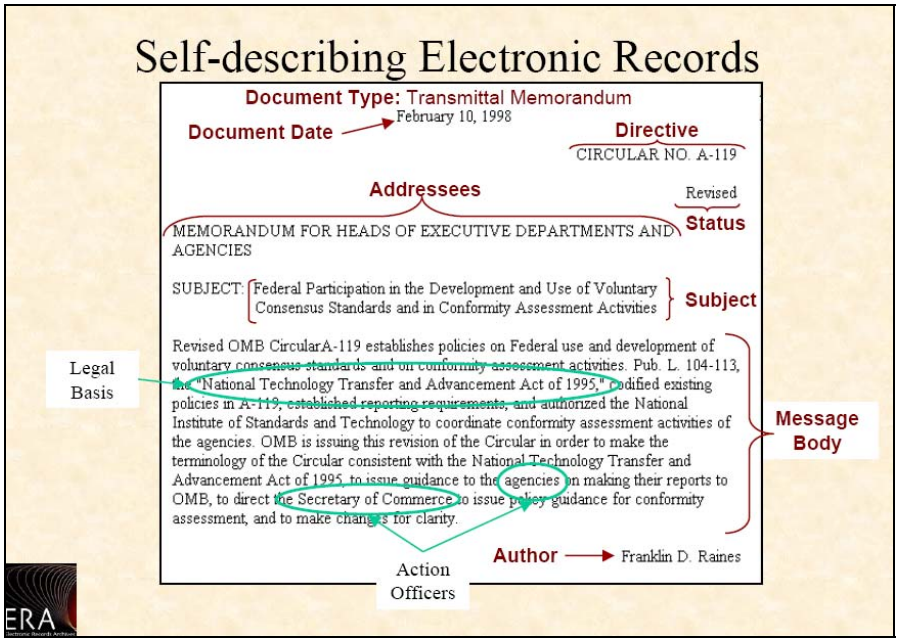


Figure 9

The programming language XML converts documents, such as the one shown above in figure 9, into separate metadata fields. When this document is stored in an electronic archive, the data fields are stored in the type of format shown below in fig 10.

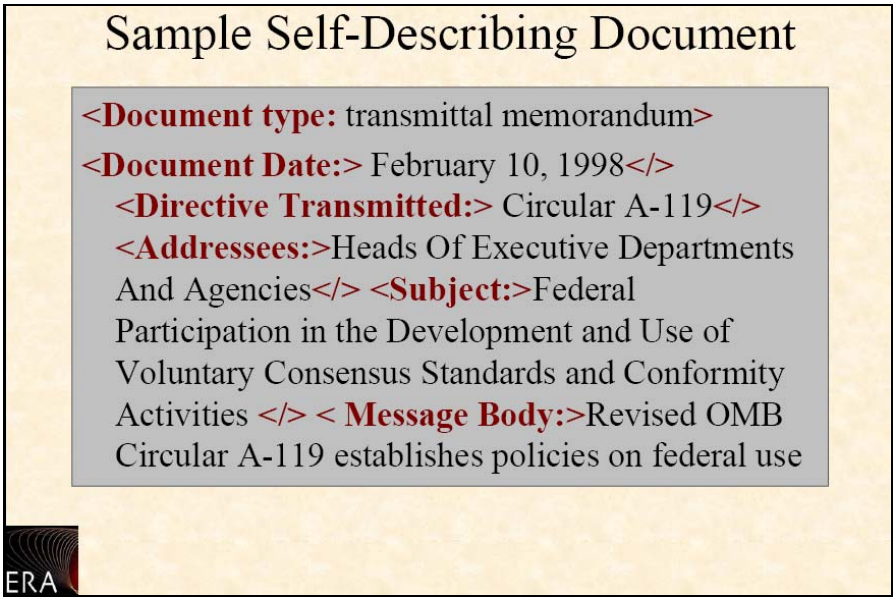


Figure 10

XML is similar to many other programming languages that use a writing system for common understanding. Foreseeing the future problems with archiving records, the International Organization for Standardization (ISO) developed a standard for adding metadata to documents called Standard Generalized Markup Language (SGML), or ISO standard 8879. As with most ISO standards the business community adopted XML, the commercial version of SGML, to address business document archiving (http://www.w3.org). XML utilizes entities, characters, character data known as "mark up", and XML processors (http://www.w3.org). An entity is a storage unit, or field of information, that holds characters. A character is, literally, each character of text within the entity. Character data are the symbols used to start and stop or provide specific information about an entity. Finally, the XML processor is the software used to interpret entities, characters, and character data.

XHTML is a version of HTML written to the more stringent XML specifications—and therefore sites that use XHTML can take advantage of the strengths of XML. One of these potent traits is the capacity for XML to withstand the rapid changes of computer technology and its potential to be viewable on hardware and with software that do not even exist yet (Cohen and Rosenzweig, 2005). Institutions cognizant of digital preservation, such as the Smithsonian, seem to be moving toward XHTML because of this flexibility. Along with Cascading Style Sheets (CSS) and a couple of more programmer-oriented technologies, XHTML is one of a handful of emerging "web standards" that professional web developers consider the basis for stable websites that will render well on different machines and that will have the greatest life expectancy (Cohen and Rosenzweig, 2005).

In 1996, the World Wide Web Consortium (W3C), which is composed of large computer software companies, established the following 10 rules for XML:

1	XML shall be straightforwardly usable over the Internet.
2	XML shall support a wide variety of applications.
3	XML shall be compatible with SGML.
4	It shall be easy to write programs which process XML documents.
5	Optional features in XML are to be kept to the absolute minimum, ideally zero.
6	XML documents should be human-legible and reasonably clear.
7	The XML design should be prepared quickly.
8	The design of XML shall be formal and concise.
9	XML documents shall be easy to create.
10	Terseness in XML markup is of minimal importance.

Table 4: XML Rules, (W3C, 2006)

Although ISO looked ahead at the advancing technology problems, neither government nor industry adopted a single standard of XML for metadata. This resulted in many different versions filling the gap for industry and government: Predictive Model Markup Language (PMML), New Mexico District Court XML Interface (XCI), LegalXML, XMLGov, Extensible Business Reporting Language (XBRL, formerly XFRML), Victorian Electronic Records Strategy (VERS), and Multivalent Document (MVD), (http://www.archives.gov/era). When an actual XML standard is finally adopted by all participants in the record archiving arena, there will still be the problem on reconciling the myriad versions to a single standard. This inevitable situation adds more difficulty to an already technically complex conundrum.

Universal Preservation Format

The Universal Preservation Format (UPF) was proposed by David MacCarn at the WGBH Education Foundation in Boston. The UPF is designed to reduce the confusion caused by the "veritable explosion of formats" (MacCarn, 2000). It also "specifies that machine independent algorithms be encapsulated within the stored media. Two strategies, the Bento Specification and the Open Media Format "are media technologies that approach the UPF concept" (MacCarn, 2000). The major disadvantage of using a single format for storing all digital information is that "no computer technical standards have yet shown any likelihood of lasting forever—indeed most have become completely obsolete within a couple of software generations" (Rothenberg, 1998). The Time Capsule

File System, proposed by Zuzga, is a similar approach to the Universal Preservation Format. It specifies a format that is "very similar to the RFC-822 format used for electronic mail" (Zuzga, 1995). However, it suffers from the same drawback as the UPF in that no single standard is likely to adapt to technologies developed in twenty years and beyond.

The Digital Rosetta Stone

A study proposed by Steven Robertson (1996) and a following Delphi study conducted by Don Kelley (2001) of the Digital Rosetta Stone showed that there is a realization by many organizations that digital information storage is beginning to become unmanageable. They also are beginning to see the urgency of seeking a solution (Jackson, 2006) due to the eventuality of obsolescence in media, hardware, and format lifecycles (Quick and Maxwell, 2005).

> *Information stored in digital format is at risk of disintegrating over a period of time much shorter than documents on preserved paper. Imagine a scenario about a CD-ROM found in an attic in the year 2045. A note is attached regarding the contents as the key to a fortune. It is assumed that compact disc media is obsolete and has never been seen by the grandchildren except in old movies. The probability of the grandchildren retrieving what is on the CD depends on the knowledge available to gain access to the media, the condition of the bit stream,*

interpretation, and proper readability of the document. The note is the difference between the disk being tossed or deciphered. "Despite the much-touted immortality of digital information…it is the letter that will be immediately intelligible fifty years from now, not the digital disk. The letter possesses the enviable quality of being readable with no machinery, tools, or special knowledge-other than that of English…" To start with, the bit stream has to be retrieved from the disc which will require a device to read it. Assuming that data is still on the disk and a reading device with the proper driver can be found to recover the data, it is necessary to use the correct software needed to decipher the format along with the information. (Rothenberg, 1999)

Robertson (1996) explored the long-term access problem and suggested an approach to retrieving and interpreting data stored on obsolete media, based on capturing the metadata. He called this model the Digital Rosetta Stone (DRS). Robertson's model was conceptual in nature and did not include details of how best to implement it.

Recognizing that there are differing strategies for systemic knowledge preservation, many approaches have been adopted by a number of groups. The DRS is designed for those instances where those other strategies fail. The DRS does not attempt to address recovery of information from media that has degraded beyond the point being readable. The DRS was designed to be a last-ditch effort to recover information where the bit stream is present, but the knowledge of either how to recover the bit stream or

interpret the bit stream has been lost. Therefore, it is to be used as a digital archaeology tool—recovering information that, until now, has been beyond reach (Kelly, 2001).

DRS Components

The DRS model is composed of three major components that are necessary to access digital information stored on obsolete devices or in obsolete formats. These components are knowledge preservation, data recovery, and document reconstruction (Robertson, 1996). Developing each of these processes accurately is critical to the success of the DRS. A diagram of the DRS model is provided in Figure 11.

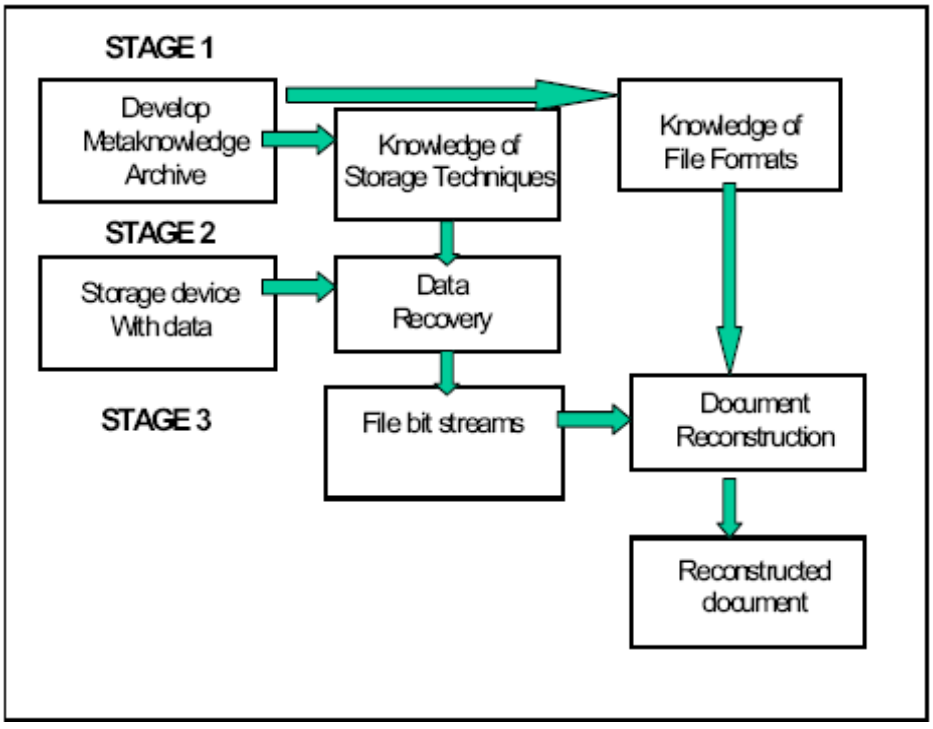


Figure 11: The DRS Model

Metaknowledge Archive

The first major process, knowledge preservation, is addressed by the Metaknowledge Archive. Robertson (1996) proposed developing a repository of information necessary to both recover the data and reconstruct the document, which he calls a Metaknowledge Archive (MKA). This archive would be created through the act of knowledge preservation and would form the foundation for the other processes of the DRS Model. In fact, without this MKA, a file stored on an obsolete medium or in an obsolete format would be completely useless, even if the bits were preserved (Zuzga, 1995; Smith, 1998). Lyman and Besser (2000) point out that we know the most about a digital object at the time we create it. We can't know for sure how we will want or need to access it in the future. Therefore, all possible information should be captured about the digital object at the time of creation. This metaknowledge describing the attributes of the digital object goes into the MKA.

Knowledge preservation is the process of collecting the information on the data storage and formatting techniques used by the designers and builders of information storage and processing devices (Kelly, 2001). This includes the technical aspects of what constitutes a bit of information on this device, how it is arranged on the device, and how it is accessed. Information is also collected from systems and applications software that identifies the file structures, along with all information necessary to recover and read the stored digital document.

Recovering the Bit stream

Armed with the knowledge of storage techniques, recovery technicians can begin restoring digital documents. Data recovery is the process of retrieving the bit stream from the outdated and obsolete medium and moving it to a current storage device. If necessary, the information in the MKA could be used to create a new medium access device (Kelly, 2001). In fact, the access method may be altogether different than the original device used. For instance, instead of building a CD-ROM drive to recover a bit stream, the DRSO workers might use a high-resolution scan of the CD and software to interpret the image. This may help if the media is fragile and may not survive traditional data access methods.

Interpreting the Bit stream

Once the bit stream is accessible to the modern computing environment, document reconstruction can take place. This is where the bit stream—manipulated using the knowledge of formatting techniques—is displayed as the original digital information object. Depending on how well the MKA has accurately captured all of the storage and formatting information, the reconstructed document can be an exact representation of the original document.

Output

The result of going through each of the stages of the DRS would result in a recovered digital information object. Given the variety of file formats, the reconstructed object

could be an encapsulated document containing metadata or a simple ASCII-text file. This flexibility gives the DRS the ability to be a long-term solution.

Methodology

Overview

This chapter describes the research objectives and the research methodology that will be used to achieve those objectives. To begin, an exploration of this study's research objectives will be outlined. Next the context of qualitative research will be defined; and finally, a look at case study research value will be described and a multiple case study design. Also, this chapter discusses case selection and data collection methodologies used in this study.

Research Objectives

The purpose of this research is to assess the different digital preservations strategies across the three selected economic sectors of society. Determining the strengths and weaknesses of different approaches will improve the efforts made by future generations in this area. By concentrating on the efficiencies of certain organizations, it is the hope of the researcher that patterns can be discovered that aid in the design of more robust digital storage strategies.

The researcher has chosen to adopt a multiple case study design of the research. The researcher plans to conduct case studies of organizations in three areas (governmental, private, and public) to learn how they are handling long term access to digital documents. Organizational data will be collected and interview data from organizational representatives will also we collected.

Qualitative Research

Qualitative research is much different than quantitative research and appropriate for collecting open-ended data with the goal of discovering themes in the data (Cresswell, 2003), and is associated more with theory building (Leedy and Ormrod, 2001). This research "uses multiple methods that are interactive and humanistic by employing such techniques as open-ended observations, interviews, and documents" (Cresswell 2003). Also, this approach lends itself to making an interpretation of the larger meaning of the data" (Cresswell, 2003). For these reasons, the researcher chose the qualitative approach. The data gathered from this research will be useful to build theory in this topic area, and the results will be used to form recommendations towards implementing and managing digital preservation technologies.

Case Study Value

Case study research has been an important tool for business researchers in part because it draws conclusions from a variety of facts and pieces of information (Cooper and Emory, 1995). The qualitative case study research is appropriate for providing description, testing theory, or generating theory (Miles and Huberman, 2002), and answers the "how" and "why" questions (Yin, 1984) by analyzing a specific case or cases in depth. The case study approach has been chosen because of the richness of the data collection aspects of this process. "This type of detail is secured from multiple sources of information. "It allows evidence to be verified and avoids missing data" (Cooper and Emory, 1995). Also because of its inherent flexibility in the design and execution of the

research, case studies are able to accommodate differences in data collection (McDonnell, Myfanwy, and Read, 2000). However, despite its strengths, the single case study has one major weakness—its narrow focus that threatens generalizability (Patton, 2002:583). To combat this threat, a multiple case study design would be adopted for this research.

The Multiple Case Study

The multiple case study design, or comparative case method, combines the findings of several independent case studies and comparisons to develop "underlying themes and other patterns" shown through an analysis of results (Leedy and Ormrod, 2001). The use of multiple cases offers potentially greater generalizability than a single case study (Ellram, 1996). Comparative, qualitative case study analysis is the appropriate and valid research design for this research because the "purpose of this report was not to portray any single case, but to synthesize lessons from all cases, organized around key topics" (McDonnell, Myfanwy, and Read, 2000).

A major disadvantage in multiple case study analysis is the time and money required to conduct quality research. A tremendous amount of effort must be expended towards identifying and selecting potential candidates, conducting and transcribing interviews, and analyzing and interpreting results. However, the benefits of multiple case studies research—if done properly—far outweigh its disadvantages. "The richness of the data obtained through the adoption of multiple perspectives is without doubt the strength of this method" (McDonnell, Myfanwy, and Read, 2000). By comparing case study

analyses, researchers may "…reveal that happenings in one case are not wholly idiosyncratic, but that there are commonalities across cases once the researcher can get beyond the specific local contextual variations" (Ritchie, 2001 referencing Miles and Huberman, 1994). Also, "the evidence from multiple cases is often considered more compelling, and the overall study is therefore regarded as being more robust" (Yin, 1984). Through deliberate consideration, it has been determined that qualitative multiple case study design is appropriate for this research. Now, the proper cases must be selected for the study.

Case Selection

Case selection is very important to a study's relevance and generalizability across other cases. Cooper and Emroy (1995) stated, "The ultimate test of a sample design is how well it represents the characteristics of the population it purports to represent." Additionally, Ellram (1996) stated that the cases of interest should have "…boundaries of interest, such as an organization, a particular industry, or a particular type of operation." Following this vain of reasoning, the researcher selected organizations from separate segments of society, each corresponding to a certain role in the economy. Darke, Shanks, and Broadbent (1998) state that "there is no ideal number of cases" to study, and they referenced Eisenhardt (1989), who suggested studying between four and ten cases. For this multiple case study, seven subjects were selected that are representative of three disparate sectors of society. Each organization has an interest in digital preservation. The organizations are listed in Table 5.

CASE	ORGANIZATION	ECONOMIC INTEREST
1	Lextron, Inc.	Private Industry/For profit
2	Teradata	Private Industry/For profit
3	EMC, Inc.	Private Industry/For profit
4	Library of Congress	Government
5	Air Force Historical Research Agency	Government
6	National Public Radio	Public Sector/ Non-profit
7	Wright State University Library	Public Sector/ Non-profit

Table 5: Case Selection

Ellram (1996) stated, "Multiple case designs should be used to either predict similar results among replications, or to show contrasting results, but for predictable, explainable reasons." To this end, these organizations were selected in the hope that significant trends and patterns would be discovered that lead to an answer of the researchers question.

Protocol Approval

Prior to conducting any research, measures were taken to protect interview subjects, execute an ethical investigation per USAF research standards and guidelines. The researcher received clearance from the Institutional Review Board (IRB). These protocol reviews and approval processes acted as external validations by outside sources which determined that the research was proper in design and ethical in practice. Copper and Emory (1995) stated, "The goal of ethics in research is to ensure that no one is harmed or suffers adverse consequences from research activities." All interviewees were

told that interviews were completely volunteer basis only. Now that the study and protocol was developed and approved, data collection could begin.

Data Collection

Now that the research design has been established and the cases have been selected and analyzed, the data must be collected. Data was collected using semi-structured interviews from a representative of each organization. Several interviews, such as the Air Force Historical Research Agency (AFHRA) and Wright-State University, were conducted over the phone or via e-mail due to large geographical separations, and some interviews were conducted in person, such as with EMC and Teradata. Once transcribed, the interviews were sent to each interviewee for review.

Interviewing

Collecting data through interviewing is a fundamental source of information for case study research (Patton, 2002; Yin, 1984), and as previously stated, this research utilizes semi-structured interviews to gather data. Cresswell (2003) states some advantages of interviewing are that it allows participants to provide historical data and allows the researcher to control the questioning. He also mentions that some disadvantages of interviews are that they provide a filtered view of the situation, the researcher's presence may bias the response, and some people may not be able to properly articulate the situation. Patton (2002) lists other limitations of interviewing such as "personal bias, anxiety, politics, and simple lack of awareness". But according to

Cooper and Emory (1995) "if the interview is carried off successfully, it is an excellent data collection technique." Darke, Shanks, and Broadbent (1998) state, "if the research area is particularly relevant to an organization and the specific research question is one which the organization needs or wishes to address, then it is more likely that they will provide access to their people and resources." The representatives that the researcher met with were all excited about being interviewed, and appeared to be interested in the results of this analysis.

Investigative Questions to the Research Questions

The execution of the research was divided into two distinct phases. The first phase comprised of gathering background information of the seven selectees and assigning them to one of the three identified categories. Phase two of the research was conducting the interviews and gathering information on the selectees.

Each organization was provided four investigative questions and their responses were recorded. Figure 12 shows the investigative questions.

IQ1) What is your organization currently doing to prepare for long-term access to digital media?

IQ2) What does your organization consider to be the "way ahead" in ensuring access to digital records?

IQ3) Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

IQ4) Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

Figure 12: Interview questions

Results of the interviews were compared and contrasted against organizational documents regarding preservation strategies and technologies. Due to the proprietary interests of a for-profit company such as Teradata, details regarding specific digital storage techniques were limited. The other organizations in the study provided ample documentation describing their respective storage processes.

Setting Boundaries for the Interviews

There are potential hazards of conducting too many interviews and asking too many questions during the interviews, thus, costing too much money and taking too much time to transcribe. The interviewees were asked open-ended questions based on the subjects answers. The open-ended nature of the interview has both its advantages and

disadvantages. It provides a rich data source, but can also provide extraneous data that can otherwise be burdensome to the researcher. In order to conduct an effective and efficient research investigation, it is imperative not only to select representative case studies, but also to select the right individuals within each case to interview. The individuals selected for interview were subject matter experts (SME).

Subject Matter Experts

One of the benefits that the qualitative researcher enjoys is "to purposefully select participants or sites (or documents or visual material) that will best help the researcher understand the problem and the research question (Cresswell, 2003)." Throughout this research, many SMEs in many organizations participated in the interviews and provided key responses to the Investigative Questions. Some of SME positions that were interviewed during this study were senior librarians, IT specialists, and archivists.

Protocol for Recording Information

Prior to every interview, informed consent was provided by each interviewee, as each subject was informed about the research and its objectives and that the interview would be recorded and transcribed. During the interview phase of the research, interviews were asked to participate in the research only if they chose to do so. Leedy and Ormrod (2001), state, "Research participants should be told the nature of the study to be conducted and be given the choice of either participating or not participating" and

"any participation in a study should be strictly voluntary." Once the interviews were accomplished, they were transcribed by the researcher verbatim.

Pilot Study

An initial pilot study was conducted after the development of the investigation tools and proposed methodology. The investigative questions were administered to a computer networking professional familiar with digital preservation. As previously mentioned, qualitative case study research design and data collection are flexible and must be able to adapt to changes or conditions in the field. Yin (1984) states, "The pilot case study helps investigators to refine their data collection plans with respect to both the content of the data and the procedures to be followed." Cooper and Emory (1995) also advocate the use of the pilot study and state that one is used "to detect weaknesses in design and instrumentation." Cresswell (2003) adds, "The research questions may change and be refined as the inquirer learns what to ask and to whom it should be asked."

Resulting from this pilot study, the researcher made several changes and revisions to the investigative questions of the research instrument. When re-administered to the professional, the new question set proved to be more robust.

Data Analysis

The qualitative narrative found in Chapters 4 and 5 resulting from this research will be an objective account across the multiple case studies. This narrative will provide

emerging patterns, conclusions, and recommendations. Once the data was collected, it had to be analyzed.

The data will be gathered, analyzed, and reported. Cresswell (2003) states, "case study research involves a detailed description of the setting or individuals, followed by an analysis of the data for themes or issues." The research presented in this study closely follows this assertion.

Content Analysis

There was a large amount of textual data to sort through after the interviews were executed and transcribed. The data analysis approach chosen for this research was content analysis. Cooper and Emory (1995) state, "content analysis measures the semantic content or the 'what' aspect of a message. Its breadth makes it a flexible and wide-ranging tool that may be used as a methodology or as a problem-specific technique." Part of the content analysis is categorizing the data and then executing pattern matching and developing common themes with frequency analysis counts from the data for each case study, across the case studies, and for the accumulated responses of all cases. Leedy and Ormrod (2001) state, "The data and their interpretations are scrutinized for underlying themes and other patterns that characterize the case more broadly than a single piece of information." After pattern matching has been executed, "an overall portrait of the case is constructed. Conclusions are drawn that may have implications beyond the specific case that has been studied" (Leedy and Ormrod, 2001).

Pattern Matching

Cresswell (2003) defines pattern matching theories as generalizations that "represent interconnected thoughts or parts linked to a whole." In this study, the "interconnected thoughts" are the opinions, best practices, and lessons learned from the cases in the research. It is the researcher's goal to link commonalities to form conclusions and answers to the question of this study. Yin (1984) states that the ultimate goal of data analysis is "to treat the evidence fairly" and to "produce compelling analytic conclusions." All transcribed data was analyzed through content analysis and pattern matching, and themes emerged from the entire data set. Now that the researcher's role has been established, validity and reliability have to be discussed more in-depth.

Validity and Reliability

Validity and reliability are critical to conducting research that produces quality results. Validity is a strength of qualitative research, and as Cresswell (2003) verifies, "is used to suggest determining whether the findings are accurate from the standpoint of the researcher, the participant, or the readers of an account." There are three different types of validity that will be addressed in this research—external validity, internal validity, and construct validity.

The first type of validity that is addressed in this research is external validity, or transferability. Cooper and Emory (1995) state, "The external validity of research findings refers to their ability to be generalized across persons, settings, and times." Transferability is the term used for more naturalistic studies, such as this one, and it asks

the question, “are there similarities between the original study and its context and any other settings where this conclusion is possible?” (Isaac and Michael, 1997). External validity and transferability are important to establish because it allows the researcher to go the next step—to draw conclusions and make recommendations for the various organizations that are concerned with digital preservation. Leedy and Ormrod (2001) report, “The external validity of a research study is the extent to which its results apply to situations beyond the study itself—in other words, the extent to which the conclusions drawn can be generalized to other contexts.” They go on to say that conducting research in a real life setting and the use of a representative example are techniques in which a researcher can employ to “enhance the external validity of a research project” (Leedy & Ormrod, 1985). This research was conducted in a real-life setting and the sample of SMEs was a truly representative sample.

Internal Validity and Credibility

Internal validity and credibility are the second type of validity that is addressed. Internal validity is “the ability of a research instrument to measure what it is purported to measure” (Cooper and Emory, 1995). Leedy and Ormrod (2001) define internal validity as “the extent to which its design and the data that it yields allow the researcher to draw accurate conclusions about cause-and-effect and other relationships within the data.” Credibility is the criteria used in naturalistic studies and asks the question, “Will the methodology and its conduct produce findings that are believable and convincing?”

(Isaac and Michael, 1997). The pilot studies conducted was a check on the internal validity of the interview instrument with respect to the overall Research Question.

Construct Validity

The third type of validity that is addressed is construct validity. Yin (1984) defines construct validity as "establishing correct operational measures for the concepts being studied." Ellram (1996) also states that using "multiple data sources," establishing a logical flow and "chain of evidence," and having "key informants review the overall case study" is important to increase research construct validity. All three of these recommendations have been included in the research design and will be addressed later in this chapter.

Reliability and Dependability

The final aspect of this research that must be addressed is reliability which "is a contributor to validity and is necessary but not sufficient condition for validity" (Cooper and Emory, 1995). Yin (1984) defines reliability as "demonstrating that the operations of a study can be repeated, with the same results." Dependability is the criteria used in naturalist studies and asks the question, "If it were done over again, would one arrive at essentially the same findings and conclusions?" (Isaac and Michael, 1997). The methodology section of this research is detailed and documented in a way so that the process can be repeated. This action ensures reliability and dependability.

Furthermore, many other steps were taken within this research in order to protect the reliability and dependability of its results. Ellram (1996) states that the two keys of reliability for a qualitative case study are the use of a case study protocol and the development of a database. This study adopted both techniques. Cooper and Emory (1995) report, "One can improve reliability if external sources of variation are minimized and the conditions under which the measurement occurs are standardized."

Strategies to Determine Validity/Reliability

Now that validity and reliability have been defined and strategies have been outlined, it is time to discuss more in depth the plan that was utilized in this research. Trustworthiness is what this researcher sought to achieve in the extrapolation and analysis of the data and in the culmination of this research into conclusions and recommendations.

The first step that was taken in this research to assure reliability and validity was conducting multiple case studies at once. Kervin (1992) states, "The validity of statistical conclusions is generally greater with a larger number of cases, and in particular small number of cases or observations can provide only very tentative conclusions." Triangulation was another technique used to assure reliability and validity. Cresswell (2003) states, "Triangulate different data sources of information by examining evidence from the sources and using it to build a coherent justification for themes." Data was collected through multiple interviews across several DOD organizations and Air Force Program Offices.

Next, member-checking was executed to assure reliability and validity of the research. Cresswell (2003) reports, "Use member-checking to determine the accuracy of the qualitative findings through taking the final report or specific descriptions or themes back to participants and determining whether these participants feel that they are accurate." First, the interviewees received the opportunity to review the transcript from the respective interview and provide feedback. Second, the data and its subsequent analysis was member-checked using two of the researcher's committee members with PhDs.

Then, the interviews were transcribed verbatim and all the data that was collected and analyzed was presented in this research—even data that did not fall into a category or went against the findings. Cresswell (2003) states, "Also present negative or discrepant information that runs counter to the themes," as this adds credibility to the research. Finally, after the pilot test was conducted, the interview questions were revised and standardized, which means each interviewee was subjected to the same investigative question set. Patton (2002) states this standardization of questions is mandatory because "how a question is worded and asked affects how the interviewee responds."

There was a massive effort made in this research to build validity and reliability much like the research conducted by Knipper (2003). He states, "It is the goal of my research to build a bridge from each validity type. The effect is cumulative. Attempts are made to minimize validity threats in sequence."

Conclusions and Recommendations of the Study

After analyzing the data and developing themes across all of the case studies, conclusions were drawn and recommendations were made based on the research. The recommendations put forth can help data preservation organizations develop their own strategies to make a more effective preservation program. Cresswell (2003) states, "A final step in data analysis involves making an interpretation or meaning of the data."

Summary of Selected Approach

Ellram (1996:102) states, "Multiple cases represent replications that allow for development of rich, theoretical framework." This qualitative multiple case study research will attempt to create theory by drawing conclusions on the current state of digital preservation across society and by providing recommendations using insights gained from the multiple interviews and emerging themes stemming from those interviews. Furthermore, using the techniques of content analysis, pattern matching, and triangulation, this researcher's findings will prove to be valid and reliable.

The remainder of this thesis is organized as follows. Chapter IV presents the analysis of the data gained from the interviews, and Chapter V draws conclusions, makes recommendations, summarizes limitations of the study, and recommends areas for future research.

IV. Case Study Results and Analysis

Chapter Overview

Chapter IV analyzes the four investigative questions (IQ) asked of all Subject Matter Experts (SME) case study groups. The answers of the SMEs are examined independently first for patterns and themes, and then the case studies are merged to produce an overall result for the corresponding sector of economic interest. For each Investigative Question, the findings of the interview and content analysis of the data are presented. Specific examples from the interviews are provided as supporting rationale. Further discussion, conclusions, and recommendations are presented in Chapter V.

Figure 13 provides a list of the investigative questions asked of all selectees.

IQ1) What is your organization currently doing to prepare for long-term access to digital media?
IQ2) What does your organization consider to be the "way ahead" in ensuring access to digital records?
IQ3) Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?
IQ4) Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

Figure 13: Interview questions

Subject Analysis

Case One: Lextrion, Inc.

Lextron, Inc. is a veterinarian supply company that operates in the central and western regions of the United States. The company was originally incorporated under the name of Great Plains Chemical Company, Inc. in Greeley, Colorado on May 16, 1967, and the founders were Park L. Loughlin and Robert C. Hummel who remain the major stockholders today.

During the 1970's the Company continued to grow and expand acquiring Elco Veterinary Supply and Parker Livestock Supply. In 1978, the Micro Tech Machine Division was opened and Lextron introduced its private label products. The Company continued its growth in the 1980's by acquiring Brawley Cattle of Dumas, Texas and Life Services of Florida; and in 1988, the Company entered the veterinary channel with a start-up company, American Veterinary Supply in San Antonio, Texas, and the acquisition of Intermountain Veterinary Supply in Denver, Colorado in 1989. The 1990's brought about more growth as a result of the ConAgra Animal Health acquisition when the ProVet Companies located in Indianapolis, Indiana; Kansas City, Missouri; Oklahoma City, Oklahoma; and Seattle, Washington, joined the veterinary side of the business. WindRiver Animal Health of Greeley, Colorado and Donnell Ag Products of Graham, Texas were also part of the ConAgra acquisition. A centralized warehouse and an outbound TeleSales Department were opened in North Kansas City, Missouri. Larson Distributing Co. in South Dakota, Poultry Health Services in Florida and Georgia and Desert Dairy Services were also acquired in the 90's. In 1994, Aspen Veterinary Resources, Ltd. replaced the Lextron label on the private label products, and today the

Cooper's Best label is also used as a result of the Poultry Health acquisition. In the 2000's, the C.W.T. organization in Gainesville, Georgia was acquired and the Company opened Lextron Dairy Services in Dublin, Texas after purchasing the assets of Dairy Tex. Lextron also became a part owner of Desert Dairy Services in Chandler, Arizona.

The various separate Corporations were consolidated into one corporation and the name was changed to Lextron, Inc. The various companies have continued to operate as Divisions of the parent Company (http://www.lextron-inc.com/History.asp).

Precision Logistics is an integrated management distribution company located in Liberty, Missouri. Precision Logistics was established in the fall of 2002 to provide logistics, warehousing and redistribution of the company's branded products (http://precisionlogistics.lextron-inc.com/). Lextron Information Systems provides Animal Management System Solutions specifically geared to meet the demanding needs of the Animal Production Industry. The company's goal is to provide livestock producers the right products and management programs to make their job easier, more satisfying and more profitable. Towards this end, Lextron Information Systems offers state of the art hardware and software components to capture Individual, Group Animal Health Information and Feed Management Data. This information in turn allows customers, veterinarians and nutritionist data analysis thru various reporting tools. The software is designed to interface with many of the current Feedlot Financial Management Systems, Benchmarking and Data Warehouse Programs, Cow / Calf Programs. Data input devices such as Chute Scales, EID Tag Readers, Temperature Probes, Feed Mill and Truck Systems are also supported (http://www.lextroninformationsystems.com/).

Responses

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

The typical electronic data record stored by this company is either a file format created/read by Microsoft Office or Informix IDS. Microsoft and Informix have developed backward compatibility into each generation of software. As a result, as long as an archive document/file was opened in the recent past, all documents are converted then stored in a currently supported file format. I would expect this method of changing file formats will be executed for the foreseeable future. We have on occasion retained old hardware running old data sets to ensure that we have old data, for example sales transaction, as far back as 10 years. As data storage has dropped in cost (hardware), it has become easier to accommodate long data retention times and lots of data. The days of limiting users to a 100MB email box for example have past and we now have some boxes greater than 2GB in size, which a few years ago would have been unthinkable. In summary two things, first, we use conversion tools to convert archive data to new file formats as necessary and second, storage is cheap compared to worrying about how much data is being stored that the company just keeps adding storage capacity.

IO2: What does your organization consider to be the "way ahead" in ensuring access to digital records?

I would expect to continue operating by converting archive files to new file formats as they become available and routinely adding storage capacity.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

Maintaining cost effective data migration paths.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

Not at this time. The cost to create/buy such a tool for a problem that largely does not exist today in most corporate environments seems out of step with corporate capital spending requirements.

Content Analysis

The Lextron organization is a company operating in the Private sector of the economy. As a "for-profit" institution, this company operates according to the bottom line of its corporate budget. Finances, Time, and Resources are the driving factors of the business operations of this company. As shown through the responses, cost appears to be a major influence concerning digital preservation at this company. Migrating to the successive iteration of the operating systems software and adding storage capacity as necessary seems to be the solutions arrived at this company. This is the result of both the dramatic reduction of

storage costs and recent vast increases of storage capacity available on the market today.

Case Two: Teradata

Teradata Corporation is the global leader in data warehousing and analytic technologies champions a slogan that they "make smart companies smarter" (http://www.teradata.com/company-overview). Teradata provides the most knowledgeable and experienced consulting professionals, highest performing technology, industry-leading innovation, and a world-class network of customers and partners to make faster, smarter decisions that give our customers a decided competitive advantage. Teradata concentrates its services to a myriad of industries including: Communications, media and entertainment, financial services, government and public services, insurance, health care and pharmaceuticals, manufacturing, retail and hospitality, and travel and transportation logistics.

Teradata offers a robust and innovative portfolio of horizontal and industry analytical applications such as data warehousing, customer management, finance and performance management, and demand and supply chain. They are able to deliver integrated, enterprise-scale analytical solutions based on the most powerful, scalable and reliable technology platform in the industry. Teradata incorporates analytical intelligence deeper into operational execution, enhances process efficiency and transforms corporate culture. As a result, the world's largest and busiest data management environments are currently running and growing on Teradata systems. The company currently has over

5,500 associates in over 40 countries and a strong diversified client base of over 850 corporate customers worldwide.

Teradata continues to retain No. 1 Leadership Position in Data Warehouse Server Evaluation Model based on the Gartner Server Evaluation Model (SEM), as published on June 18, 2007 (http://www.teradata.com/company-overview).

Responses

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

We maintain a current platform of technology. If I have got everything from 1990 to the end of 1999 on disk storage on generation-a type of disk, and I decide to go through my data center and upgrade to generation-b disk; I migrate all that data to gen-b and then I start adding my new stuff to that gen-b. I don't need that gen-a data anymore because I have migrated the data to my new platform and I have access to it. So in the IT business, we migrate that key information from platform to platform, then duplicate it, and then eliminate the old. It is different from archiving. Say 10 years from now I have this data on gen-a disk. I am on gen-z now, how do I read it? We don't typically do that. We migrate the data all the way through, and retain it that way, rather than go back to that old data. I would argue what we have seen is that it is cheaper to do that. Disk is getting so cheap. Where we used to buy disk that was so big for a dollar, we now can buy that same amount of space for a quarter. So its cheaper to migrate that data than trying to put that data out in a DRS.

IQ2: What does your organization consider to be the "way ahead" in ensuring access to digital records?

Part of the concern I had is technology is changing so rapidly. As soon as you come up with a solution for your technology today, its going to change. So how do you make sure that whatever you've done today fits the new technology? Technology is moving at light speed in some areas. I still have a lot of media on 5 ¼ disks. I have 5 personal computers and none of them have a 5 ¼ reader. So I have all of these floppies sitting in a box at home but can't do anything with them. You hate to throw them away. But what can you do with them? Now we are in the same situation with the 3 ½ disks, first with the standard format, then the high density. We don't have readers for those anymore, but I still have drawers filled with them. So you go to the PC world with the Microsoft DOS. I have all kinds of DOS applications and disks with DOS applications. Then you go into Windows then Windows 2003. Its changing so rapidly, and Vista is out there now. We haven't incorporated Vista yet, but that's the problem we have. Again, the problem is changing so fast that whatever you implement today could be outdated tomorrow. We haven't cracked that nut yet.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

Through Y2K, not only did we go back through all of our systems and make sure they would work on Jan 2; but we also started to think about data

retention. How do we do that? We came up with what we called the Greenfield process. Which is exactly what you are talking about here? The Greenfield process is how do we take data and Archive that data so whenever we need it in the future we will have the ability to use it. Whether it is on tape or disk or however it is written, how can we put it in a format so we can use it in the future? Conceptually it was a big project and it had a big push. But it never took off. Because one, we could not figure out how to do it. Two, it was a lot of work.

In most cases, the DRS model, as I read it and understand it, is more pertinent to us from an n minus one or two generations standpoint. I might have to go back one or two versions of that technology or hardware or media, but not anything further than that. In most cases, it is extremely important when going from technology of today to technology of tomorrow—that migration. Almost every vendor I go to will have that migration plan already in place to support n-1 or n-2 versions. MS Word is a great example. If I am sending you a document and you are in Vista and I am 2 versions back, can you see it? Probably. If you send me a document, can I look at it being 2 generations back? Probably not. Now everybody and their brother are out there making translators and that fuels transition. That is more important to us than maintaining historical stuff past a point in time or older than that. We still have reels of tapes in the data center that we have no reader for. We have the tape, though. Another example, our tape comes in cassettes as you can imagine, but now they're bigger. Some might have stated as 100 gig tapes, now we are up to 700 gigs. The format is slightly

different because its more dense as they have figured out how to get more data on the media. But it comes down to hardware. What hardware do you have that can read those things? In our environment here, we don't upgrade our hardware often enough. If we did upgrade it every 18 months or 2 years, it would become more of an issue to be able to maintain those tapes and pull it off. Again as we migrate stuff forward, how critical is that old stuff if we have all the new stuff? So our migration process is very key to us there. So while it is an issue and concern, we haven't got the pressure or corporate government requirements to drive us to find a solution, yet.

To throw some numbers out there, Oracle is one of the largest enterprise resource applications out there. Just one application for Teradata has over 3000 tables and each table could have hundreds of fields within it. There are literally terabytes and terabytes of data with primary pointers and secondary pointers linking all of this data together. So if I go into customer A, it could be just an overwhelming amount of data tied to that customer. What we did instead of going to the Greenfield approach and trying to dump all this data was to back this [data] up to tape. At that time across the 50, we had a standard platform, so we took a server and set it aside. We decided we are going to keep this server and keep this tape drive and set it in a corner and hope it doesn't die on us. So as we went to Australia to Japan and all of these countries and migrated their data into our newer systems, we maintained this rack of tapes and this one system sitting over there. So for any reason, a government inquiry or lawsuit, if we had to go

back and look for the information it would be there. We had this old technology and it is complete off warranty, no maintenance, no spare parts available. Its sitting over there and nobody [is allowed to] touch it and if you do you will get shot. That was the option we chose to go with. So until somebody says, "Nope, we don't need that any more," that is our solution to this problem because we did not have the time or money or resources for a Greenfield or Rosetta stone approach.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

Part of the problem, again, is cost and resources. Where does it stack in a company's strategic or operational plan? Traditionally it doesn't get over the line. It is not something seen as a critical issue as of now. Should it be? I don't know. It hasn't bitten us yet so you wonder how critical it is.

In our business its not as critical because outside of that statutory requirement, 7 years, there is no other requirement we have. It is not like we are the Smithsonian trying to preserve history. We are trying to keep business running and make money and move forward and maintain that 7 year back set of data for legal issues: taxes, law suits, etc. Once it is 8 years old, it has no significant value to us. We are not doing it to maintain our corporate history. You look at the annual report and you can go back any year and get the high level

numbers but not a line by line item description, its nuts. No we don't have a critical aspect or need to maintain that history past that time.

But if you look at corporate America and ask for something 10 years ago, it is probably going to be contained within their annual report. If it is not, I don't think anyone would come and ask a specific question about our installed customer base 10 years ago—what sites did Customer A have, where were those sites located, what were their annuity programs for those sites. Again, that's old history we don't maintain. It isn't pertinent to our business moving forward. It's a little different from tracking history. I think its different from a historical perspective compared to a business perspective. Now from a media standpoint, we all have the same problems, same issues. How do I keep track of my daughter's pictures when she was one? How do I save those and how do I pull those up now? That's probably more important than our business requirement. I am setting up a picture book for my daughter's graduation right now. How do I format and pull up that data? That is a media issue and I don't think we have those same issues in business.

Content Analysis

Teradata is also a for-profit company. The representative interviewed was very forth-right about operating procedures at the company. Although the DRS model promises value from a historical perspective, it has little use for this company's preservation needs.

This is due to the limited retention time that Teradata holds on to their customer's data. Because of legal constraints instituted by the federal government—Sarbains-Oxley Act, companies are accountable for documentations pertaining to their electronic records from the past five to seven years. This seven year window is the time reference Teradata operates within. By and large, Teradata is not concerned with maintaining data beyond the legal requirement.

This is for several reasons. One, it is not cost effective for Teradata to ho beyond what is required. To do so would put a drain on revenues due to the increased operating costs involved in such an undertaking. Second, they have not received any pressure or direction from upper management to seek a solution to a problem that might not affect them; and third, Teradata has been given no standardization guidance from regulatory agencies in how to proceed with digital preservation effort.

Migrating to the next generation of technology is what Teradata practices as a preservation strategy. Keeping their technology platform current is of the utmost importance to the company.

The Teradata representative also added that standardization is necessary for any type of preservation effort. Without common practices and procedures in place, there is little chance of compatibility of data that is produced by different organizations. The representative suggested that some organization needs to take the lead of a standardization movement.

Case Three: EMC, Inc.

The EMC Corporation is a manufacturer of software and systems for information management and data storage. The corporate headquarters is located in Hopkinton, Massachusetts. EMC offers a wide range of products aimed at enterprise storage. These include things like hardware disk arrays and storage management software. Its flagship product is Symmetrix, which is the foundation of storage networks in many large data centers around the globe.

When EMC began in 1979, the company started out as only a manufacturer of memory boards, but the company has steadily evolved since then into a leading data storage provider. During 2006, we invested nearly $4 billion in Research & Development and strategic acquisitions that strengthened our core capabilities and extended our reach into new, rapidly growing markets (http://www.emc.com/about/emc-at-glance/corporate-profile/index.htm). Through a series of acquisitions and partnerships, EMC has become the largest provider of data storage platforms in the world (http://en.wikipedia.org/wiki/EMC_Corporation). Its major competitors are IBM, Hewlett-Packard, Hitachi Data Systems and Network Appliance.

EMC considers information to be a business's most important asset. The company's focus is to develop systems that provide tools that can help businesses capitalize on a comprehensive information infrastructure. These infrastructures are the versatile foundations on which organizations can implement their information lifecycle strategies, secure their critical information assets, leverage their content for competitive advantage, and automate their data center operations. With an information infrastructure in place, businesses can avoid the potentially serious risks and reduce the significant costs

associated with managing information, while fully exploiting its value for business advantage (http://www.emc.com/about/emc-at-glance/corporate-profile/index.htm).

EMC has a long tradition of innovation and leadership. This commitment led to IDC's designation of EMC as a market leader in the external storage systems, total storage software, and virtualization software markets. According to the Gartner Magic Quadrants, we lead the industry in enterprise content management, midrange enterprise disk arrays, storage resource management, security and information and event management, Web access management, and storage services. We hold the most stringent quality management certification from the International Organization for Standardization (ISO 9001), and our manufacturing operations hold an MRP II Class A certification (http://www.emc.com/about/emc-at-glance/corporate-profile/index.htm).

Responses

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

It is one of those things as I was reading the DRS. I am in a presales role so I have some sales objectives. When I am talking about an EMC product for, example Cetera, its specific market-nitch is archiving. That is all it is. It is a storage device, a digital storage device, but it is not intended to be used in day to day operations. It is strictly there for a repository. Long term, also. Now when we are out selling technology, long term is 5 – 7 years. One of the things I noticed in the paper there were some periods that were specified up to 70 years and some with no specification at all. Its

an area I haven't given much thought to. On one hand the concept seems reasonable…to an extent. One of the things I was thinking about. Let's go out 100 years and somebody is doing an excavation somewhere and they uncover a disk drive. There is probably some information on there. How am I going to get it off? First of all let's say there was a label and you can read it and it says it's a Western-Digital and that tells you something. On WD disk drive, information was stored on 512 byte sectors and there is x-number of tracks and cylinders and all that other technical mumbo jumbo that is all well and good and necessary information, but at the end of the day, how do I talk to the disk drive? How do I interface with that device? If I don't have the electrical interface as part of that DRS model then it does me no good to know how the data is stored on that medium itself because I don't have any way to get it off. That is just one of those thoughts that crept into my mind.

IQ2: What does your organization consider to be the "way ahead" in ensuring access to digital records?

That is a good question. Obviously you have some experience with Teradata. We got set up their infra structure. In that case it was a straight forward migration. There wasn't any preservation we had to worry about. In the case of Lexis and other service type bureaus, companies that provide computer based technology services of some sort, they have to deal more with those types of problems that you were identifying like keeping track of reel to reel tape drive. People graduating now don't even know what those are now. In a reel tape drive and I can't buy parts anymore for it. That is

what I started out doing in this business, working on those things. They went from 9 to 18 to 36 tracks and a lot of other technical mumbo jumbo that went in them, but the point is quite valid, and especially in a service type environment, I suspect that as time goes on, there's going to be data reclamation types of businesses that say okay you've got a 9 track tape and I've got a tape drive that will help you recover that data for you. But at some point, if that is my business, I'm not going to be able to buy parts to fix my 9 track tape drive. So some point in time that tape is going to die.

I will be honest with you, as I thought about this problem… probably every sales outfit is not really thinking about this problem in terms of more than 7 years, that is pretty much the standard, the legally, obliged time you have to keep records. You're really not going to be concerned much more beyond that. There are going to be some exceptions. NARA would be one, obviously, but from a business perspective, most are only concerned with the 7 years I have to keep this stuff. As a sales force, frankly we are not worried either. 25, 50 years from now that is somebody else's problem. That is the prevailing attitude. I would bet money on it. Now as I thought about it, it becomes an archiving problem vs. a day to day data processing business problem. Teradata processes data day in and day out. Oracle financial databases and financial applications they have to process every day. My opinion is you wouldn't worry so much about the digital preservation of the data you have to process day to day but as you archive your information ,your quarterly statements and annual reports, whatever regulatory requirements that Teradata has to deal with, you are going to archive some of that data. That is the data, or type of data, I think as an industry we need to start thinking about. Call it an archiving API. Basically

it is a standardized interface. Take Microsoft for example. The difference between Windows and MAC is with Macintosh, Apple is very proprietary. They write the code, they test the code, and it works great, yadda, yadda yaddah. Microsoft took the open approach and has what they call a hardware abstraction. If you think of your hardware as memory and disk drives as your lowest layer; Microsoft lays a hardware abstraction layer on top of that. One side talks to the physical devices and knows those devices; the application side of that layer could care less about what kind of device it is. It talks to a standardized interface. So then a read request going to a disk drive, whether its a WD or Sea gate or whatever, the read request going to the a digital disk drive is the same because of the standardized interface. We offer an archiving product, but it is just a bunch of disk drives. You want to store your annual records for 7 years? Dump them in here and we will keep them for 7 years and you can read and retrieve them as you wish, assuming you have the application to do so. But from a storage and repository standpoint, that will remain constant. Over those seven years, EMC will be coming out with a new technology, you know last year's model had 250 gig. Next year's model will have 500. Next year's model will have terabyte drives. We will sell them a new unit that will have more capacity and have the same footprint, and we will migrate the data off of there and onto here. From a preservation perspective, that really is nothing more than a migrating from n – 1 technology to the current generation technology. I wouldn't categorize that as a preservation initiative because the data is still digital and in the same format. The application is still there that can access the data. But if all archived data was archived in a consistent fashion whether it was stored on a Cetara or a Sun

disk drive; if it is an archived piece of data, there should be a standardized interface. Then with information that was stored in 50 years from now, I know that it was stored this way and all I have to do is pop in this interface. The underlying hardware technologies are going to change.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

We had a conversation with NCR about this subject. It was called Greenfield. That was the issue they were struggling with. Today we could probably export the data out of an Oracle database and then import it into XML. That makes it pretty transportable from a data perspective. OK…but what am I going to do with it. Just because its transportable doesn't mean its useable. I have got to have the application that created it to make sense out of it. There are a myriad of issues for the long term data technology evolution that are really mind boggling.
Certainly, businesses are probably more discreet in their data retention policies. Especially after the last several years as these new regulatory requirements came out. Sarbanes, Hippa, FCC related stuff. Depending on the agency, there are lots of different regulations on how long you have to keep your data. If you are involved in a law suit you have to be able to provide certain types of data. To say, "I don't have it." is no longer an excuse. Companies have become very sensitive to keeping lots of data, but on the other hand they also have to be sensitive to getting rid of that data the moment that it

is no longer legally required to be kept because if you have the data it can be used against you.

I think from a business standpoint, most have IT staffs that deal with technology day in and day out, and I believe they are savvy enough to make a firm decision to get rid of data that is older than x, whatever that time is. My first comment would be your point on cheap storage. It is so easy anymore to just buy more storage and keep doing what we have been doing. But it does reach a point where even with today's storage technology that are untold orders of magnitude better than even 10 years ago; meaning that you can store a terabyte of information in the past and it would fill up this whole building, but now that amount of information can sit in the corner and you won't even know its there.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

To answer your question, no. But I think the rest of the answer is that most companies have an IT infrastructure. It has grown with their business. They have the means and technology both hardware and software and people to be able to preserve their data, from day to day, month to month, year to year. Back to the earlier comments, the private sector is not concerned with the really long term like 50 – 100 year time frame. They are concerned with the 2 – 7 year time frame. It is pretty easy to keep yourself current over that period. Because being in the storage technology side of this business, we see most of our customers keep their assets 3 – 4 years. So every 3 – 4 years they are bringing in the next generation replacing the n-1 technology. That is just

basic migration. I don't see a lot of focus on it in the private sector. Because to them its not a problem. They are rolling the assets so frequently. No one anymore is going to come out with a technology that leaves everything behind. It happened once. It was before my time, but as I remember IBM basically left behind all the previous generation stuff. There was such a cry from the business community that they vowed never to do that again. That was one of those early learning lessons, lessons learned. You can't do that. People are not going to leave behind the last 50 years of doing stuff. It was such a dramatic change in the hardware architecture that it was incompatible with the current technology. As a business EMC can't afford to alienate our customers so we provide technologies so that you can migrate from one technology to the next. Now NARA is going to become more and more digitized. Those are the places where these types of topics would be more relevant because you have some form of data that was created on paper 50 years ago. We now have scanned it and made an image of it. What we need to do now is make sure we can read that image in the future. There almost becomes a division in the discussion of the data. The day to day data that businesses run on is one branch of the data storage retention tree, if you will. The longer term requirements and archiving is going to be another branch of that tree. And they're separate requirements, really. I need high performance over here so I can run my business day to day, but once I've used that data and its reached the end of its useful life but I need to keep it around for whatever, I archive it. And I put in over on the archive medium. Having a standardized archive interface or archive abstraction layer so that, going forward, it

doesn't matter what kind of data is archived, the process is always the same. I am not aware of anything going on in that area.

Content Analysis

The representative from EMC was very informative, not only about EMC, but also offered insight to the IT profession in general. EMC was the third organization interviewed that operates with in the private sector.

The subject's views on the relevance or usefulness of the DRS model fell in line with previous interviewee's opinions. In the private sector, companies are only going to be interested in maintaining their data for as long as they have to. This usually falls within the range of five to seven years. According to this representative, it would not make sense or be financially justifiable to any organization to go beyond that limit. Consequently, the DRS model holds little value in the private sector, in the representative's opinion.

However, he acknowledged that due to the advances in technology and the proliferation of digital devices and media, such a preservation system like the DRS may very well become a necessity. As more and more consumers immerse themselves in digital data that is collected through cameras, recorders, music players, etc., a DRS device will be needed.

To illustrate the point, the representative brought up the example of photographs. He no longer uses an analog film camera—all the cameras he uses now are digital. However, he has shoeboxes full of photos taken with film cameras. He can easily pass

these items down to his children and grandchildren because they are able to be handled physically. The digital photos he has stored on hard drives will be more problematic to be passed down to successive generations due to the medium they are stored on. He would like to pass them along, but keeping them in the current format so the future additions to the family line may be able to view them could prove difficult.

The representative's photo dilemma is not an isolated one. More and more families are experiencing similar difficulties. The problem is exasperated through the proliferation of digital media in society. Where will all the data that is collected and going to be collected be stored? Also, what guarantee do we have that the formats of today will be compatible with the technology of tomorrow?

Case Four: Air Force Historical Research Agency

The Air Force Historical Research Agency (AFHRA) is the repository for Air Force historical documents. The Agency's collection began in Washington, DC, during World War II. It moved in 1949 to Maxwell Air Force Base, where it is still located and is an adjunct site of Air University. AFHRA's mission is to provide research facilities for professional military education students, the faculty, visiting scholars, and the general public. Its collections consist of over 70,000,000 pages devoted to the history of the service, and represent the world's largest and most valuable organized collection of documents on US military aviation (http://afhra.maxwell.af.mil/about.asp). AFHRA's collection consists of two broad categories of materials: Unit Histories and Special Collections.

Unit Histories

The major portion of the collection consists of unit histories that the various Air Force organizations have prepared and submitted periodically since the establishment of the Air Force History Program in 1942. Reporting requirements have changed from time to time over the years, and the submissions vary in quality (http://afhra.maxwell.af.mil/holdings.asp). When viewed in unison, the unit histories, with their supporting documents, create a clear picture of Air Force activities.

Special Collections

The coverage provided by unit histories is supplemented by special collections, including historical monographs and studies; oral history interview transcripts; End-Of-Tour Reports; personal papers of retired general officers and other Air Force personnel; reference materials on the early period of military aviation; course materials of the Air Corps Tactical School of the 1920s and 1930s; working documents of various joint and combined commands; miscellaneous documents or collections of various organizations, including the US Army, British Air Ministry, and the German Air Force; USAF individual aircraft record cards, and a large collection of material relating to the USAF activities in the war in Southeast Asia and Operations Desert Shield and Desert Storm. The Agency accessions approximately 2,000,000 pages of historical material each year, including the annual and quarterly histories of Air Force units as well as additions to the special collections. Working closely with the Air Force Historian and the History Offices of the major commands, the Agency conducts an oral history program to record important historical data that would otherwise be lost. The Agency also gives special attention to

the acquisition of personal papers of value for documenting Air Force and airpower history.

Over the years, the Agency's collection has been used by the Air Force for preparation of plans, development of programs, analyses and evaluations of operations, staff studies on many subjects, textbooks and other course materials for Air Force schools, student papers and theses, orientation and indoctrination of personnel, and many other purposes. The collection has provided information regarding military aviation in general and the US Air Force in particular to Congress, the military services, and other government agencies. The collection has been used extensively by scholars, students, and writers, for books and monographs, master's theses, doctoral dissertations, magazine articles, and TV and movie scripts (http://afhra.maxwell.af.mil/holdings.asp).

Responses

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

We are currently digitizing our collections. In terms of what we are doing for future long term preservation is a little more problematic. Currently we are scanning our documents into a PDF format so that they can be stored in that format, which at the moment seems to be one of the more universal formats that can be used and read by almost anything. And of course we back up those all the time. The other thing we do is we produce microfilm. We take those analog paper copies and scan them making digital copies. Then we create analog microfilm from the digital images. That kind of seems

backwards, maybe. But it is really the only way, in my mind, to make sure you can see it. All you need is the film and a light source. From an archival perspective that is something we do with historically significant documents.

Another thing we do is when we scan things into PDF; we also shoot TIFF images which are lost-less images. That means when we have to pull them up and manipulate them, all the information is there and the bottom doesn't fall out, resolution wise.

IQ2: What does your organization consider to be the "way ahead" in ensuring access to digital records?

One of the things you mentioned about the DRS that seems to be key in all of these is there has to be some sort of standard, acceptable format that we can use for these digital images. You mentioned Apple vs. Microsoft. Well that situation is kind of merging. When you look at Intel based MACs anymore, you can run any kind of those programs and it will run just about any image from either system. So that part is kind of merging in a sense, but the real challenge I think is a standard format that can be universally read for archival purposes.

Well one of the drivers in all of this should well be NARA. But I can tell you from meeting with them several times and working with them often from the AF publications world that the Records Management program is broke. That is kind of a side issue. But the point is that the NARA has for years side stepped this issue about digitization. For one thing it is so huge; it was just another format headache for them. They have recently

come around, though; and their plan is to make sure they have a viable back up system so that they can access theses things. Of course the NARA problem is a hundred fold more complicated than what we have to deal with.

Our dilemma is the breadth of our collection. You will see if you talk to any university around the country, some will do rare books; some will do other things like art work or printed material. Here at the AFHRA, I have any kind of video you can think of—Betamax, VHS, 16mm film, 8mm film, slides. I have all of these different kinds of things that can be digitized. We have spreadsheets, databases, etc. And all of those things have to be migrated at some point. The problem I have with other media is I have to migrate it before I lose the capacity to run it. Try to find a super 8 projector today. NASA has millions of bytes of data that they can't access anymore because they were originally recorded on wire recorders. The machine no longer exists. That is a problem and part of that goes back to standardization. How do we get all of this stuff into one format so we can read them?

For some time I was interested in XML as something we can use. I am just familiar enough with it to be stupid and dangerous; but you need some type of tagging and flagging convention to build a metaknowledge archive that you can reinterpret. If you don't have the metadata, there is nothing to interpret. You really need the metadata that relates to the document itself in terms of what it is, but also how do you build a metaknowledge archive of the software and those bits. That is another very complicated problem. If I understand what you are saying, after you record that and you interpret the bitstream, then, like DNA, you recreate the animal based on what is in the metadata.

Metadata is one of the absolutely critical things for us when we store these documents. Because if I put them in a 4 terabyte server, we expect the computer in essence…Well let me back up a little.

There is a couple of ways we can do this. We could build the file structure so that it looks like paper that we are very comfortable with; but I will tell you that only after a little while, the files become quite cumbersome because I am doing it right now. The file structure becomes almost unwieldy. Because if we start talking about current contingency operations, I am getting these things that are getting huge—there are hundreds of files in there. So the other solution, which is probably the way we are going to go, is to assign a number to the document which will be a key field in a database. That number will correspond to that document and all the supporting documents that go with it. Then we create a link that not only brings up the metadata that we develop but brings up the document as well.

That puts a lot of faith in the processors on the server and your database that they are going to be able to find that document. The key to make that work is the metadata that is built when you get intellectual and physical control of the document. We get histories, for example. When we archive those histories we look for key words that are important and the other important things in the document—all of the document's administrative markings that go into the archive. All of the electronic histories that we get now and all of the supporting documents and citations that are linked back to the document are done in MS Word. The Air Force has decided that Windows Office will be the universal software package used in their systems. We have those Word docs, but I am

not sure what the solution is for them. How do you maintain those Word doc's and ensure they will be available 10, 15 years from now? I think what you are talking about developing a metaknowledge of that software program. Are there universal keys that we can draw out and piece together? It makes perfect sense to me, but I don't have a clue about how you would do that.

It is really in developing the metadata. Not only the information that is in the documents but also how it relates to the code that is written for the software that essentially runs the document. You have metadata that relates to the physical attributes and the intellectual attributes of the document. But what you are talking about here is the metadata that shows how to map the software applications that it runs on.

So you have to have two things to get a duck out of the data when you clone it. First you have to know it is a duck. Second you have to know the gene code that makes it a duck. So it seems to me that sort of thing is doable. But we have to be careful about saying all we have to do is just…we have a tendency to say "just". The devil is in everything after "just".

It is a fascinating idea. What AFHRA does now is that we back up everything extensively. We try to have all the applications to run the programs. And as an archivist, I have to have all the equipment to run the archives. Sometimes that involves using legacy equipment.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

When I first got assigned here twenty-five years ago, I got working on taking WWII documents and digitizing them. The idea was to send the paper to NARA or something. To make this long story short, over the twenty-five years, to make the idea work, the technology was never there to do it. All we could so is make a long electronic list of metadata and put it into a database. We had no way of digitizing those documents. It wasn't anything more than an electronic card catalog. Well about 10 years ago, we started working on phase two of this project, which is to digitize the document and then create the doc-link so you could not only call up the metadata, but also link to the document and cross reference it, etc. etc.

But even then, the technology had not caught up to the vision we had. We have come to the point now that we have all of the software technology to make those links. We can make them almost on the fly. We are currently working on a work flow management system. When electronic documents are sent to us, they are automatically assigned the next key number so it can go directly into the server. We then can build metadata on the document and we can process it all the way through the system linking metadata information to it as needed—who touched it, where it is located, etc. That workflow system is the accumulation of a twenty-five year process that started out by creating an electronic card catalog. As the technological innovations came on board, we decided that we could use that for digitalizing. We have scanners that can digitize 400 pages a minute, and we automatically convert paper to digital images, to TIFF images, to PDFs, but none of that was available in the beginning. Every time we took a step, a new

door opened with new challenges, new problems, and other doors. So as we went through this, it was kind of like taking two steps forward and a half step back.

What I am saying is…What I hear you saying is absolutely wonderful, and we must solve this problem or we will have nothing more than a huge pile of mush of 1s and 0s that we can't do anything with. Or we will have a few high-priests of the archives that will keep a small amount of this stuff, but in actuality it won't be of much value to anyone because of the lack of context.

When Beale AFB was downsizing, I received some hard drives from them. I had no way of knowing what the application was, I don't know what operating system it used, I could not read them. We got rid of them eventually because they were of no use to us. The folks at Beale couldn't read them either. There might have been the answer to world hunger stored on them, but I couldn't get to it. There is such a huge, vast amount of data. When I talk about the AF records program, it is amazing. I can't imagine how they sell that.

But as you know, when you start working with electrons and 1s and 0s, if you don't start working with an organized structure up front, you end up with garbage at the end. The biggest thing is how do you organize it? That is exactly what he is talking about when he talks about a metaknowledge archive. How is it organized? Then how to I interpret it and recreate it? But it seems to me what the next piece is someone needs to work on the mathematics side and computer side of this problem to show we can do this. I can't imagine why you couldn't. It is amazing the things we are getting knowledge of now and doing now. If we can think about quantum mechanics and things being in two

places at the same time, then this should be kid stuff. I think for a couple of reasons it hasn't been done yet is this problem is like world hunger. It is such a big problem that everyone throws up their hands and says I can't get my arms around this. The reality is if you did it on a small scale first and find out that you could do this part, then it would lead to other things. You have to take it one bite at a time.

I have gone around and around with IT folks about some of this. They were visionary about how electrons moved around but they had no idea how a researcher worked. And I wasn't confident enough with the computers and systems to say this pile of mush belongs to the paper and know what the data means. As I learned more, I found out the important thing is the metadata. It is describing the documents. Whether you use XML to tag these things or whatever, it is really building the model first. You can't recreate it at the end if you don't have all the information. When I talk about physical attributes, I mean what is important in the document—the administrative markings, the classification. How will we put it in our big pile of stuff here so that it will make sense? You have to build that on every document. Electronic documents give us the ability to do key work searches. If you build that carefully then I don't have to build an abstract on the article. I can just do a key word search. It is all about building it well upfront.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

The operative word is "practical". I think theoretically it is absolutely useful. When you are taking about digital archives, it is absolutely something that would be key.

Just in general terms, if I am talking about what the DRS is, it is essential because we will never be able to get our arms around all the babble that is being created. So this idea is essential. But to say it is a practical way? That implies there is something there to be useful, and then I can say if it is practical or not. Theoretically I can say that it is the way to delve into. I am not sure if the mechanics have been fleshed out enough to do so.

I think theoretically it is right on the money. The devil is can we do these things. What are the essential programming codes and do they work everywhere? What are those and is there enough? I hate to keep using DNA, but is there enough DNA and genes there to rebuild the thing. So there is some very fundamental technical research that has to be done to say this thing works. Conceptually, it is brilliant. I am surprised I didn't think of it. It is amazing when you start taking about the interpretation of the bitstream. When you start thinking about 1s and 0s and when you start thinking about fiber optics, its just either on or off. So there is a big leap between that simple idea and then interpreting and translating it to reality. I think this is a great idea, but one of the things that strikes me about what you are doing is…I work with IT people a lot. I am an archivist. My job, as I see it sometimes, is to translate. That involves translating what historians want into technical information and dialogue techno-geeks understand. I can speak enough of that stuff also to be dangerous. My point is, you almost have to be a liaison between the techno-geeks that have to do the mathematics to say we can do this. You have to take those important items you want to keep track of and compare that against the mathematics work to do it. This is a wonderful idea but it is kind of like

antimatter or dark matter. We kind of know something is there, but we are not sure what it exactly is and if we can ever use it. The first question is, is any of this doable? Is it theoretically doable? Is it physically doable? Well if the answer is yes, then you point down the path and show what the research steps are to answer these questions from the collection of metaknowledge. Then identify what exactly is that metaknowledge and start building this DRS. If the math doesn't work, though, then you are just talking about this nice theory you revisit every couple of years.

Content Analysis

The interview with the representative from the AFHRA also provided several interesting findings to the researcher.

This representative definitely sees a need for some type of future preservation strategy. Whether this is to done through some type of programming language is unsure, but the metadata would be absolutely essential in developing any kind of preservation system.

He also advocated the continual migration of materials to the next viable form of format. As the representative put it, "it is the only way to be sure." By converting documents to "mainstream" formats, AFHRA's archives are given a greater chance at survival. Also, the representative mentioned that in addition to the conversion to the next format, a microfiche of the document is created. In this case, all you would need is the film and a light source.

He suggested that some organization—NARA—should try to form some type of standardized format.

Case Five: Library of Congress

The Library of Congress is the *de facto* national library of the United States and the research arm of the United States Congress. Located in Washington, D.C., it is the largest by shelf space and one of the most important libraries in the world. Its collections include more than 30 million catalogued books and other print materials in 470 languages; more than 58 million manuscripts; the largest rare book collection in North America, including a Gutenberg Bible (one of only four perfect vellum copies known to exist); over 1 million US Government publications; 1 million issues of world newspapers spanning the past three centuries; 33,000 bound newspaper volumes; 500,000 microfilm reels; over 6,000 comic books titles; the world's largest collection of legal materials; films; 4.8 million maps; sheet music; and 2.7 million sound recordings. The head of the Library is the Librarian of Congress.

The Library serves as a legal repository for copyright protection and copyright registration, and as the base for the United States Copyright Office. All publishers are required to submit two copies of their copyrightable works to the Library. This is known as a mandatory deposit. Nearly 22,000 new items published in the U.S. arrive every business day at the Library. Though it does not keep all of these submissions in its permanent collection, an average of 10,000 items are added per day. When included in the permanent collection, it is the responsibility of the Library of Congress to maintain

copies of all works in the English language. This collection adds up to approximately 130 million items with 29 million books.

It is estimated if the print holdings of the Library of Congress were digitized and stored as plain text, it would constitute 17 to 20 terabytes of information. The Library currently has no plans for systematic digitization of any significant portion of its books.

However, The LOC has taken strides towards investigating the effectiveness of digital preservation. The LOC has started the American Memory project that makes millions of digital objects available at its American Memory website. American Memory is a source for public domain image resources, as well as audio, video, and archived Web content. Also, the LOC is responsible for digital preservation milestones such as the development of the National Digital Information Infrastructure and Preservation Program (NDIIPP).

The Library of Congress also provides an online archive of the proceedings of the U.S. Congress, including bill text, Congressional Record text, bill summary and status, the Congressional Record Index, and the United States Constitution (http://en.wikipedia.org/wiki/Library_of_Congress).

Responses (Note: Many of the answers offered to the researcher were given in the form of document links. The researcher has synthesized responses from the interviewee and the information contained in the linked documents referenced by the interviewee to formulate answers to the investigative questions.)

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

The Library of Congress has taken a collaborative approach to the collection and preservation of digital information in order to remain relevant and useful to Congress and its constituents in the digital age. No single institution can do the job of collecting, preserving and making available all the information in digital form that that students, teachers, researchers and lifelong learners have come to expect will be available at the touch of a mouse.

In December 2000, Congress asked the Library to lead a collaborative project, called the National Digital Information Infrastructure and Preservation Program (NDIIPP), in recognition of the importance of preserving digital content for future generations. Congress passed special legislation (Public Law 106-554) appropriating $100 million to the Library of Congress to lead this effort. The goal of the program is to develop a national strategy to collect, archive and preserve the growing amounts of digital content, especially materials that are created only in digital formats, for current and future generations.

IQ2*:* What does your organization consider to be the "way ahead" in ensuring access to digital records?

Our vision is to enable the Library to continue collecting, archiving and sustaining records of cultural knowledge and creativity; to build trusted access and preservation tools and services; to advance open standards and best practices to ensure

the sustainability of the nation's cultural records; and to engage national and international partners in preserving the world's critical digital assets.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

Because of the variety of efforts on digital preservation, however, it is equally clear that there will be no quick convergence of methods in the digital preservation community. Every system is rightly designed to fulfill the goals of the sponsoring institution, and as institutional goals differ, so do the systems. While this does not damage the goal of digital preservation (in fact, it enhances it, because heterogeneity guards against system wide failure), it also means that the trivial interoperability of "everyone uses the same tools and formats" and the deeper interoperability of "everyone uses the same conceptual model" are both unattainable, now and for the foreseeable future.

Because this sort of simple interoperability is outside our grasp, the NDIIPP architecture must support institutions that are inclined to cooperate with one another on issues of digital preservation, but who have differing technological systems in place.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

(No response)

Content Analysis

Although this was an untraditional interview that used a chat room type format, the interview with the representative from the LOC organization provided several interesting findings to the researcher and offered additional insight to preservation problems not previously considered. The representative indicated that the LOC has made strides towards coming up with a solution to preserving digital media. The LOC has formed several partnerships with other organizations to perform research in these areas. From their efforts, they have created the NPIIDS format and achieved other preservation milestones.

The representative added that much more work and research needs to be accomplished and that the DRS might prove to be a valid approach in the future.

Case Six: National Public Radio

NPR is a nationally acclaimed provider of news, information, and entertainment programming, and is the media industry leader in sound gathering and production. The world's first noncommercial, satellite-delivered radio network, NPR is an independent, private, nonprofit membership organization funded primarily through its own service-generating activities.

NPR was incorporated in 1970 pursuant to the Public Broadcasting Act of 1967, but it is not a government agency. NPR is not a radio station itself nor does it own any radio stations, but NPR programs can be heard on more than 860 public radio stations across the United States.

The mission of NPR is to work in partnership with member stations to create a more informed public — one challenged and invigorated by a deeper understanding and appreciation of events, ideas and cultures. To accomplish our mission, we produce, acquire, and distribute programming that meets the highest standards of public service in journalism and cultural expression; we represent our members in matters of their mutual interest; and we provide satellite interconnection for the entire public radio system.

NPR programming is distributed via satellite to more than 860 stations nationwide –in all 50 states, the District of Columbia, Puerto Rico and Guam. Each station designs its own format by combining local programming with offerings from NPR and other sources to best serve its particular audience.

NPR produces and distributes more than 130 hours of original programming each week, including the award-winning newsmagazines *Morning Edition*® and *All Things Considered*®; entertainment programs such as *Car Talk*; music programming such as *The Thistle & Shamrock* and *Marian McPartland's Piano Jazz*®; and a variety of talk and information programs, including *Talk of the Nation* and *News & Notes*. NPR News also produces extended coverage of special events and breaking news, from the award-winning coverage of Sept. 11 to regular election specials (http://www.npr.org/about/nprworks.html).

Responses

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

NPR maintains an extensive collection of materials dating back to over 30 years. Our collections include transcripts of texts from our programs, such as "Morning Edition" and "All Things Considered", audio files, and video files. We consider it our mission to create and maintain these products for our customers—our listeners—to document the stories broadcasted, and to enable re-use of content.

We have donated 25 years of cultural programming collections to the Library of Congress and also sent a large amount NPR archives to the University of Maryland. Both of these institutions have the capacity to properly store and maintain our donations. In addition, at the offices of NPR, we maintain a "cold room", which is used for storage purposes of our media recordings from our programming.

Currently, NPR is in planning stages to implement a digital asset management (DAM) system to facilitate desktop access to the archive. This DAM system will be used to integrate the many forms of media produced from our programs. Text, audio, video, web content, and other associative data will all be linked together and have the ability to be accessed from a single location. Hopefully this will eliminate the "stove-pipes" that each type of media resides in and bring in all together.

IQ2: What does your organization consider to be the "way ahead" in ensuring access to digital records?

Regular file migration to the next supported format.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

Knowledge preservation seems like a useful first step toward any preservation strategy. All of the elements, taken separately or together would advance preservation efforts in theory.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

I'm not sure we're yet at a place to be able to debate the merits of one strategy over another. First, there would need to be a high-level manager tasked with the technical side of preservation efforts, not just information technology and production efforts.

I'm also not sure I understand the effort involved to create the metaknowledge archive. I believe more technical research should be done in this area.

Content Analysis

The interview with the representative from the NPR organization also provided several interesting findings to the researcher.

NPR is the single greatest generator of online content presented on the Web. Most of these contributions are podcasts of the shows that are aired on NPR broadcast stations. Consequently, NPR has an enormous store of data that it has difficulty

maintaining. Some of their collection has been turned over to other organizations such as the LOC and University of Maryland.

NPR is currently looking into acquiring a digital archive management system that will enable them to store, organize, and link the different types of data that their organization processes and creates.

Case Seven: Wright State University Library

Named after the world-famous Wright brothers, Wright State University in Dayton, Ohio, continues their spirit of innovation. The university serves nearly 17,000 students, offering more than 100 undergraduate and 50 Ph.D., graduate, and professional degrees.

Wright State's impact can be felt in the areas of business, education, health care, science, information technology, and the arts, to name just a few. Our partnerships with area hospitals and health care providers help provide the highest quality of patient care, discover new cures, and train tomorrow's health care providers. Wright State showcases award-winning students and funds need-based WSU scholarships annually for Miami Valley students in the fine and performing arts. Collaboration with industry and government helps find solutions to today's problems and bolsters the local and state economy by making the Miami Valley competitive in a global marketplace.

Responses

IQ1: What is your organization currently doing to prepare for long-term access to digital media?

In terms of…this is a hard question to answer. I am looking at it from a personal perspective as the head of digital services. I am sure there are other things on campus or even in the library that other folks are doing. But what we do is when we scan anything, we scan it at a very high resolution. With the thought that we would be able to convert it and change it as technology moves along. We also store those materials on a server that is taped backed-up every night. We also have backups that are on CD-ROM that we store off site on some items that haven't made it to the server yet.

I think we are a unique case, in my department, we scan materials from the university and the special collections area and try to make it available online and like I said our mindset was we should always scan it at a very high resolution and keep it and back it up so we can use it later.

IQ2: What does your organization consider to be the "way ahead" in ensuring access to digital records?

Well that has been our stance that whatever supersedes that next technology we will always migrate and move forward our old stuff to that. It sounds easier than really doing it. When you look around at what you have you realize it is going to be tough. We recently started an institutional repository. One of the guarantees of that repository is that if you submit something like a tiff image. We will guarantee we will always migrate that and make it accessible. That is our policy through this repository. So there are

certain types of files that we have selected and decided that we are confident enough to translate those file types in the future rather than some proprietary file that is out there.

If a faculty member tries to put some oddball thing into our repository, we will take it. In will be viewable for now. But we will tell them upfront that we are not committed to it. You might not be able to see these 50 years from now. But if you submit something in a known format like a tiff or jpeg, we will move it along to the next future format.

IQ3: Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?

I think all of them really. You see I am coming from a different angle. I was trained as an archivist. I worked at the Ohio Historical Society and we received governor select papers and they were all on a format that we couldn't open. We all stood there and went "Uhhhhhh". We had to work with a competing company that helped us to get it all transferred over and made gobs of money doing it. I also worked on projects where I worked with a governor that was concerned with e-mail and how to keep track of that. I know that we receive donations now in a box with a stray floppy disk on the bottom. I imagine as time marches on it is going to be more floppies and CDs than paper that we will be pulling into the archives. If there was a way that we could simply extract them and make them useable again, that would be great.

IQ4: Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?

I would think so. I have heard of other folks and actually seen computers with software on the computers that make it seem that you would be able to go into this giant warehouse and find the right program to open your disk. That seems like… How in the world would that ever work to save all that stuff? But if there was a sort of program that could extract those materials and re-create the documents, that would be great. It would be much simpler and a much more practical thing to do. It seems possible.

We are already creating metadata, just simple descriptive metadata. We save what type of scanner we used, what software we used, and that is the type of information we use for all our images. So I think that is the type of metaknowledge you are talking about. We are kind of taking steps in that direction if this thing ever got off the ground.

Content Analysis

The interview with the representative from the WSU organization also provided several interesting findings to the researcher.

Summation of Results

This section answers the research questions using the data collected from the interviews and the content analysis and pattern matching executed in Chapter IV.

Research Question:

What is the current state of digital preservation efforts across the four major sectors of U.S. society?

This question was answered through the following investigative questions administered to the seven selectees of the case study group. All of the organizations maintained some type of data storage system and added upgrades to the system.

IQ 1) *What is your organization currently doing to prepare for long-term access to digital media?*

All of the organizations employed some type of migration strategy into their data preservation systems. Organizations belonging to the private sector upgraded their systems more frequently due to expanding needs of their business. However, it should be noted that the set of data the private companies are maintaining is much smaller and manageable than the collections other institutions are responsible for. Upgrading a smaller set of data is an easier accomplished task than upgrading a large and diverse collection. This enables the private sector to be more on the cutting edge of technology and offer their customers the latest products and services.

To prepare for long-term access to digital media, companies from the private sector add storage as needed. The organizations from the government and public sectors also add storage when required, but also are seeking other methods to linking and utilizing the stored data.

IQ 2) *What does your organization consider to be the "way ahead" in ensuring access to digital records?*

Again, the private sector's answer for this question is continual migration to the next generation software and technology. However, it was added that some type of standardization should be developed so that no matter what type of business or profession that a company belong to it will be able to operate with the same type of formats as everyone else.

IQ 3) *Are there any elements of the Digital Rosetta Stone model that would prove helpful towards digital preservation at your organization?*

From a commercial standpoint, all of the interviewees from the private sector said they would not think that the DRS model would prove beneficial to their organizations. This was because of the cost in manpower and capital investment such an undertaking would require. It would prove to be financially irresponsible.

From a private perspective, however, two of the interviews noted they see great value in a DRS type system in personal and governmental use. Personal scenarios involving photographs of children and keeping them available to future generations were expounded upon by both interviewees separately.

The governmental and public sectors both acknowledged the value in the DRS model. They agreed that the flood of different formats adds difficulty to the preservation problem. A DRS approach would make sense and would solve many of those problems.

All three aspects of the DRS were noted to be of value to the representatives from the government and public sectors.

IQ 4) *Do you consider the Digital Rosetta Stone model to be a practical solution to the preservation needs of your organization? Why or why not?*

For reasons already stated, the private sector representatives did not think the DRS would prove to be a workable solution to the preservation needs of their organizations. This was due to the costs involved with implementing such a system.

The governmental and public organizations while championing the model's value, raised concerns over the model's ability to move into reality. Some representatives questioned the feasibility of such a solution while others thought more of the mechanics of the DRS needed to be better explained and developed.

V. Discussion, Conclusions, and Recommendations

Chapter Overview

Chapter IV provided the data collection and analyzed the results of the interviews with the subject matter experts in the various case study groups. Chapter V draws conclusions from the data analysis and makes recommendations to Air Force acquisition professionals with regards to acquiring the services of contractors on the battlefield. As Patton (1990) observed:

> *It is important to understand that the interpretive explanation of qualitative analysis does not yield knowledge in the same sense as quantitative explanation. The emphasis is on illumination, understanding, and extrapolation rather than causal determination, prediction, and generalization.*

Chapter V draws conclusions, using the data analysis from the investigative questions to answer this study's overall research question. Next, implications for digital preservation efforts are highlighted as best practices and lessons learned are discussed. This study concludes with a brief discussion on the limitations of this research, recommendations for future research, and a final summary.

Discussion of Results

Patterns from the Private Sector

First, the companies in the private sector are comprised by a diverse group of specialties, across the economy. Although each company serves a specialized nitch and provides unique services to its customer base, some commonalities across the corporations emerged regarding preservation practices, as demonstrated by Lextron's response to IQ3:

> *"In summary two things, first, we use conversion tools to convert archive data to new file formats as necessary and second, storage is cheap compared to worrying about how much data is being stored that the company just keeps adding storage capacity."*

This concluded the interviews of selectees from the private sector of the case study group. After further analysis, several trends appeared from the research. One such finding appeared to be the difficulty in developing a preservation system such as the DRS. The Teradata representative pointed out several times that such a notion was proposed in the company, but eventually was discarded due to the time and financial burdens such an undertaking would necessitate.

> *"So if I have a floppy with information on it, 7 years from now how do I read it? And not only how do I read it, but how do I put it into my current system? I may have migrated from gen 1 to gen 2 to gen 3 type of technology. Or, I may have moved from company a to company b, to*

company c's product. So how do I take that information seven years later and make sense of it? We never cracked that nut. One, because we did not have the time or resources to have people to go and do that, but we called this our Greenfield approach. The thinking was if you have a blank slate or a green field kind of environment to do it however you wanted to, how would you do it so that it would work in any situation? As we went through this, the resource impact was huge. Even though everyone had projects and objectives that it applied to, it kind of trickled down off the list."

Through the examination of the data collection from the interviews, the researcher came to understand that organizations belonging to the private sector operate under the constraints of the market and business world. The "bottom line" is the driving factor of decisions affecting the business and its profitability. While the problems associated with digital preservation are not going away, the business world is doing little to contribute to finding solutions. To illustrate the point, the organizations in the private sector complied with the letter of the law found in the Sarbanes-Oxley Act. However, once information stored on their systems transgressed outside the dictated seven year window of compliance, the data was discarded.

"*If you are involved in a law suit you have to be able to provide certain types of data. To say, "I don't have it." is no longer an excuse. Companies have become very*

sensitive to keeping lots of data, but on the other hand they also have to be sensitive to getting rid of that data the moment that it is no longer legally required to be kept because if you have the data it can be used against you."

--EMC Representative

Once the time retention requirement of company's financial and other data was satisfied, virtually no company went beyond the required mandate to hold on to information and records. Financial reports and e-mails can be stored into some type of retention system that updates itself by staying within the seven year window. When information extends beyond that range, it is discarded.

Granted, most information and data collected and stored by businesses is financial in nature and not worthy to be preserved for a long time. It does make sense for those type of organizations to discard of the data that is no longer useful. But where the private sector can contribute to the digital preservation problem is in the area of research and collaboration.

As they need, companies upgrade their systems to the next generation of technology, but no evidence was presented to the researcher that any profit-seeking organization was trying to develop a long term preservation system like the DRS. This sentiment was echoed by all of the representatives from this section of the case group. Due to the dramatic drop in cost of digital storage, businesses find it hard to justify creating a new preservation standard.

"I would argue what we have seen is that it is cheaper to do that than try to figure out that other problem. Disk is getting so cheap and where we used to buy disk that was this big for a dollar we now can buy that same amount of space for a quarter. So its cheaper to migrate that data than trying to put that data out in a Greenfield of DRS."

—Teradata Representative

"It is so easy anymore to just buy more storage and keep doing what we have been doing."

--EMC Representive

Patterns from the Government Sector

Because of the purpose of governmental institutions, their approaches to preserving posterity and public records goes far beyond the five to seven year requirement practiced by the companies of the private sector.

Research was recommended by the representatives from this group into formalized practices and procedures. In the public and governmental sectors, organizations are keenly aware of the pitfalls of digital preservation and are aggressively seeking solutions and trying to develop systems that will handle the preservation needs of the future.

"…we must solve this problem or we will have nothing more than a huge pile of mush of 1s and 0s that we can't do anything with. Or we will have a few high-priests of the archives that will keep a small amount of this stuff, but in actuality it won't be of much value to anyone because of the lack of context."

--AFHRA Representative

The Library of Congress has accomplished a tremendous amount of research in this area. However, the dissemination of their research has been limited, in the researcher's opinion. The representatives from each section had little to no knowledge of preservation efforts outside of their jurisdiction. Perhaps this was due to a lack of preparedness on the interviewee's part or of the spontaneous nature of the interview itself, but absence of knowledge about what other organizations are doing or what they are interested in was manifest.

The NDIIPP system created and developed by the LOC is very similar in design and purpose to the DRS model first developed by Steven Robertson. Both models use a central type of repository for the stores of metadata and document information. Based on this finding, the researcher may conclude that the DRS model has significant merit through the LOC's usage of the DRS model's preservation methods.

Pattern from the Public Sector

The preservation efforts in the public sector mirror the concerns held in the government sector and are deeply invested in the area of digital preservation. Both

representatives interviewed from this section stated that their organization tries to maintain collections of data from a myriad of formats.

Limitations of the Research

Case study design which has certain limitations when attempting to generalize conclusions and recommendations. The SMEs that were interviewed all had different experience levels which affected the data. The organizations that were selected all were involved with different preservation policies and systems. Furthermore, the case study group was not comprised of equal amounts of SMEs, and each organization served a different mission with a different focus with respect to digital preservation efforts. Although intentional, this definitely had an effect on the outcomes of this research, as the patterns analysis showed clear delineations along economic sector lines.

Conclusions & Recommendations

The researcher would recommend that more collaboration and information sharing is needed between the public and government sectors of the economy. With a greater body of knowledge to draw from, more options become available. Many of the problems that the representative from the Air Force Historical Research Agency was encumbered with could be alleviated with programs and technology developed by the LOC.

Developing and implementing data format standards may extend the time between migration iterations. Organizations may be able to conserve scarce resources (time and

money) by advocating for and adopting standards. Well-defined and universally-accepted standards implemented today may enable people to read digital information in the future without having to rely on the particular system that created the information. Outside the library and archival communities there has been little recognition of the importance of long-term preservation of electronic information. This gap remains in place even though many private companies are now entering the digital arena. Companies such as Teradata and EMC have stepped in to provide long term preservation solutions to businesses. However, the business world's definition of "long term" is in actuality rather short—seven years.

We are quickly approaching the technical ability to capture and store any and all information. This would solve some archivists' uncertainty in regards as to what to preserve. But a heavy-handed approach such as this would undoubtedly lead to other problems. By capturing everything from the momentous to the trivial, how do you categorize events and documents appropriately? You might very well be creating your own hay stack when trying to find your needle.

There is a critical need for future research in preservation field, not only for the military, but across the entire economic spectrum. Future research in this area could shorten the knowledge gap in this area and focus on specific preservation problems, potentially creating an ironclad policy and preservation plan.

Bibliography

Baru, C., Moore, R., Rajasekar, A., & Wan, M. (1998). The SDSC Storage Resource Broker. In S. A. MacKay (Ed)., *Proceedings of CASCON '98 Conference*, *Nov. 30–Dec. 3, 1998, Toronto, Canada* (p. 5). Toronto, Ont.: IMB Canada.

Bearman, David. "Toward a Reference Model for Business Acceptable Communication", 1994, available at:<http://www.lis.pitt.edu/~nhprc/prog6-5.html>

Bennet, J.C., "A Framework of Data Types and Formats, and Issues Affecting the Long Term Preservation of Digital Materials", 1997. Available at: <http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/rept011.pdf>

Bray, Tim; Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau (September 2006). Extensible Markup Language (XML) 1.0 (Fourth Edition) - Origin and Goals. World Wide Web Consortium. Retrieved on October 29, 2006.

Briggs, Robert O., Vreede, Gert-Jan De, Nunamaker, Jay and Sprague, Ralph. "Special Section: Context Driven Information Access and Deployment" Journal of Management Information Systems, Vol 21 No. 4 Spring 2005 pp. 7-9.

Butler, Meredith A., Issues and Challenges of Archiving and Storing Digital Information: Preserving the Past for Future Scholars, Journal of Library Administration, Hawthorn Press, Page Range: 61 – 79, 1997.

Cacas, Rita, ERA Infopaper, <http://www.nara.gov/era>.

Carlin, John W. "Records, Records Everywhere, But How Are They Going To Survive?," The Record. September, 1998.

Cohen, Daniel J. and Roy Rosenzweig, Digital History: Preserving Digital History—A Guide to Gathering, Preserving, and Presenting the Past on the Web, UNIVERSITY OF PENNSYLVANIA PRESS, 2005.

Cooper, Donald R. and C. William Emory. *Business Research Methods* (5th Edition). Chicago, IL: Irwin, 1995.

Conway, Paul. "Digitizing Preservation." *Library Journal*, (February 1, 1994), 42–45.

Creswell, John W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA: Sage, 2003.

Darke, Peta, Graemen Shanks, and Marianne Broadbent. "Successfully Completing Case Study Research: Combining Rigour, Relevance and Pragmatism," *Information Systems Journal*, 273-289 (1998).

Dollar, C. *Authentic electronic records: strategies for long-term access,* Chicago, Il. Cohasset Associates, 2002.

Eisenhardt, Kathleen M., Building Theories From Case Study Research, *Academy of Management. The Academy of Management Review;* Oct 1989; 14, 4; ABI/INFORM Global, pg. 532.

Ellram, Lisa M. "The Use of the Case Study Method In Logistics Research," *Journal of Business Logistics*, 17(2): 93 – 138 (1996).

Gantz, John F., "The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010", IDC White Paper - sponsored by EMC, March 2007.

General Devices – Unlocking the Data in EMS Data Transmission – Rosetta – Unlocking the Data in EMS Data Transmission – http://www.general-devices.com/7100b.htm (11/26/2006).

Goh, Chris – The Northern Rivers Echo – http://www.echonews.com/703/chris_goh.html

Granger, Stewart. "Emulation as a Digital Preservation Strategy," D-Lib Magazine, October 2000, Volume 6, Number 10.

Gupta, Amarnath and Richard Marciano (2001). Methodologies for the Long-Term Preservation of and Access to Software-Dependent Electronic Records, <www.sdsc.edu/NARA>.

Hedstrom, 2003, Margaret, IT'S ABOUT TIME: Research Challenges in Digital Archiving and Long-term Preservation, WORKSHOP ON RESEARCH CHALLENGES IN DIGITAL ARCHIVING AND LONG-TERM PRESERVATION, Aug 2003. http://chnm.gmu.edu/digitalhistory/links/pdf/preserving/8_4b.pdf

Robertson and Heminger, 1996, Alan R. and Steven B. Robertson. The Digital Rosetta Stone: A Model For Maintaining Long-Term Access To Static Digital Documents, *Communications of the Association for Information Systems*, Vol 3 No 2 January 2000.

Huberman, Michael A. and Matthew B. Miles. *The Qualitative Researcher's Companion*. Thousand Oaks, CA: Sage, 2002.

Ireland, 1998, Jonathon B., Assessing Air Force Major Command Awareness of and Plans for Dealing with the Impact of Technological Obsolescence on Long-Term Storage and Retrieval of Digital Information, 1998.

Jackson, William, "Modern Relics: NIST and Others Work on How to Preserve Data for Later Use," *Government Computer News,* June 19, 2006.

Kelley, Don M. A Delphi Assessment of the Digital Rosetta Stone Model. MS Thesis, AFIT/FIR/ENV/01M-10. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March, 2001. (0 1172 00387449)

Kervin, J.B., Methods for Business Research, New York, Harper Collins, 1992.

Knipper, Michael E. *Determining the Value of Automation in Commercial and USAF Supplier Evaluation Systems.* MS thesis, AFIT/GAQ/ENV/03-06. Air Force Institute of Technology, Wright-Patterson AFB OH, March 2003.

Kochtanek, T. R. and K. K. Hein. "Delphi study of digital libraries," Journal of Information Processing and Management Vol. 35 Issue 3: 245-254 (May 1999).

Leedy, Paul D. and Jeanne Ellis Ormrod. *Practical Research: Planning and Design*. Columbus, OH: Merrill Prentice Hall, 2001.

Lyman, Peter and Howard Besser. "Time & Bits Managing Digital Continuity." Essay. N. pag. http://www.lesk.com/mlesk/auspres/aus.html. 10 July 2000.

MacCarn, Dave. "Toward A Universal Data Format For The Preservation Of Media." Essay. N. pag. http://info.wgbh.org/upf/papers/SMPTE_UPF_paper.html. 19 September 2000.

McBride, Sean Patrick. Data Organization and Abstraction for Distributed Intrusion Detection. S. McBride, *M.S. of Computer Science Thesis*, North Carolina State University, Raleigh, NC, 2005. http://www.lib.ncsu.edu/theses/available/etd-04052005-182228/unrestricted/etd.pdf

McDonnell, Ann, Myfanwy Hones, and Susan Read. "Practical Considerations in Case Study Research: The Relationship Between Methodology and Process," *Jounal of Advanced Nursing*, 32(2): 383 – 390 (2000).

Moore, R., Baru, C., Rajasekar, A., Ludascher, B., Marciano, R., & Wan, M., et al. (2000). Collection-based persistent digital archives—Parts 1& 2. *D-Lib Magazine, 6*(3). Retrieved January 19, 2005, from http://www.dlib.org/dlib/march00/moore/03moore-pt1.html and http://www.dlib.org/dlib/april00/moore/04moore-pt2.html.

Moore, R., "Building Preservation Environments with Data Grid Technology", American Archivist, vol. 69, no. 1, pp. 139-158, July 2006.

Moore, R., A. Rajasekar, M. Wan, "Storage Resource Broker Global Data Grids", NASA / IEEE MSST2006, Fourteenth NASA Goddard / Twenty-third IEEE Conference on Mass Storage Systems and Technologies, April 2006.

Novak, Ryan M. Going To War With Defense Contractors: A Case Study Analysis of Battlefield Acquisition. MS Thesis. AFIT/GAQ/ENV/04M-08. School of Engineering and Management, Air Force Institute of Technology (AU), Wright-Patterson AFB, OH. March 2004.

Pace, Andrew. "Digital Preservation: Everything New Is Old Again," Computers in Libraries: 55-58 (February 2000).

Patton, Michael Quinn. *Qualitative Research & Evaluation Methods (3rd Edition). Thousand Oaks, CA: Sage, 2002.*

Preserving the Past to Protect the Future *The Strategic Plan of The National Archives and Records Administration 2006–2016.*

Quick, Ken and Maxwell, Mike, "Ending Digital Obsolescence" Affiliated Computer Services, Inc. (ACS). Dallas, Texas, USA. January 20, 2005. http://microfilm.net.au/pdf/ACS%20Datasurance%20White%20Paper.pdf

Rajasekar, A., Marciano, R., & Moore, R. (1999). Collection based persistent archives. In *Proceedings of the 16th Annual IEEE Symposium on Mass Storage Systems, March 15–18, 1999, San Diego, CA* (pp. 176–84). Los Alamitos, CA: IEEE Computer Society Press.

Rajasekar, A., Wan, M., Moore, R., Kremenek, G., & Guptil, T. (2003). Data grids, collections, and grid bricks. In *Proceedings of the 20th IEEE Symposium/11th NASA Goddard Conference on Mass Storage Systems and Technologies, April 7–10, 2003, San Diego, CA* (pp. 2–9). Los Alamitos, CA: IEEE Computer Society Press.

Reagan, Brad. "The Digital Ice Age." Popular Mechanics, December 2006. http://www.popularmechanics.com/technology/industry/4201645.html

Ritchie, Jan E. "Case Series Research: A Case for Qualitative Method in Assembling Evidence," *Physiotherapy Theory and Practice* (Case Study Research). 17: 127 – 135 (2001).

Robertson, Steven B. Digital Rosetta Stone: A Conception Model For Maintaining Long-Term Access To Digital Documents. MS Thesis, AFIT/GIR/LAR/96D-8. School of Logistics and Acquisition Management, Air Force Institute of Technology(AU), Wright-Patterson AFB OH, December, 1996. (AAM 8385 0 1172 0021663)

Rothenberg, Jeff. "Ensuring the Longevity of Digital Information." Expanded version of "Ensuring the Longevity of Digital Documents" Scientific American, January 1995,(Vol. 272, Number 1, pp.42-47). Revision: February 22, 1999 http://www.clir.org/programs/otheractiv/ensuring.pdf

Schloman, Barbara. "Information Resources: Now you see it, now you don't: The ephemeral nature of digital information." Online Journal of Issues in Nursing, April 2003. http://nursingworld.org/ojin/infocol/info_11.htm

Shepard, Thom UPF Project Coordinator - "Presentation to the Musis Library Association 02-13-98" http://info.wgbh.org/upf/slides/MLAtext.html

Smith, Abby. "Preservation in the Future Tense," Council on Library and Information Resources No. 3. May/June 1998. http://www.clir.org/pubs/issues/issues03.html.

Stepanek, 1998, M. (1998, April 20). "From digits to dust [Electronic version]." *Business Week,* April 20, 1998, *3574,* 128-129.

The Strategic Information Resources Management Plan of the National Archives and Records Administration, September 1, 2006, Version 5.0, http://www.archives.gov/era

John Garrett and Donald Waters. Preserving digital information: Report of the Task Force on Archiving of Digital Information, May 1996. Accessible at http://www.rlg.org/ArchTF/.

Yin, R. K. *Case Study Research: Design and Methods* (Applied Research Method Series) (Vol 5). Newbury Park, CA: Sage, 1984.

Zuzga, Brian "Tape Archiving Using the Time Capsule File System" http://www-swiss.ai.mit.edu/~boogles/papers/tcfs-thesis/thesis.html

http://www.nara.gov/era

http://www.sdsc.edu/NARA/Publications.html

http://info.wgbh.org/upf/slides/slide01.html

http://www.xmlfiles.com/xml/xml_intro.asp Retrieved February 24, 2007.

http://jmis.bentley.edu/articles/v21_n4_p7/index.html

http://commons.wikimedia.org/wiki/Image:Moore_Law_diagram_%282004%29.png

Datawatch, "2006 E-mail Management: An Oxymoron?" AIIM Study, 10/11/2006.

"Email Archiving: Who Does What," Byte & Switch Insider, Vol. 4, No. 7, July 2006.

Sarbanes-Oxley Act of 2002, Public Law No. 107-204, 116 Statutes at Large p. 745.

http://www.networksolutions.com/learning-center/Sarbanes_Oxley_guide.pdf

Laura McLemore quotation from WGBH, "Migration," Universal Preservation Format http://web.archive.org/web/20041108161056/http://info.wgbh.org/upf/survey/survey07.html (1 of 3)6/28/2005 9:27:15 AM.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
27-03-2008	Master's Thesis	Sep 2006 – March 2008

4. TITLE AND SUBTITLE
A Multiple Case Study Analysis of Digital Preservation Techniques Across Government, Private, and Public Service Organizations

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

TSGT David P. Gough

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GIR/ENV/08-M08

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Intentionally left blank

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The process of record keeping has evolved through time. As our technology advances, so does our ability to manage information. We have progressed from paper-based records to new digital techniques and formats to store records. However, digital storage is not the "Holy Grail" answer to preservation and storage problems. Digital storage is confounded by multiple problems, also. Some of these problems are, but not limited to, lack of standardization and legal guidance, proprietary formats, and the fragility of the digital medium. This research examines several organizations that are deeply involved in digital preservation and tries to identify common practices and problems across the industry.

15. SUBJECT TERMS

Digital Preservation, Digital Rosetta Stone, Migration, Emulation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Alan R. Heminger, USAF
U	U	U	UU	136	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-7405 x7405 (alan.heminger@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18